havana

human and vertebrate analysis and annotation

Otterlace • Zmap • Blixem • Dotter user manual

Written by and contributions from

Charles Steward (cas@sanger.ac.uk)
Gemma Barson
Laurens Wilming
Ed Griffiths
James Gilbert
Jennifer Harrow

23 May 2011

Blank

Contents

Otterlace 2
Starting an Otterlace Session 2
DataSet chooser 2
Transcript chooser section 5
File menu: Manage the Otterlace editing session. 5
Subseq menu: Editing operations on the transcripts listed in the window. 6
Clone menu: Edit properties of each of the clones (one or many) opened in the otterlace editing session. 7
Tools menu: Useful things to run on the genomic sequence being annotated.. 8
Transcript editor section. 10
File menu: Saving, closing, plus windows for showing translation and selecting supporting evidence. 12
Exon menu: Tools for editing the exons. 13
Tools menu: Informative operations to run on the transcript. 14
Attributes menu: Controlled annotation vocabulary for transcript and locus.. 15
Quality control 16
Zmap 18
Opening Zmap 18
Main Zmap interface. 19
Navigating in Zmap and zooming options. 20
The Focus Feature vs the Marked Feature 22
General Zmap display features 24
Functionality of the features at the top of the Zmap display. 26
Show feature details 28
Exporting features for gene objects 29
Bumping features 28
Searching for a sequence in Zmap 30
Searching for a feature in Zmap 31
Selecting single or multiple features and hiding/showing them 33
Rapid variant construction 34
Splitting windows in Zmap 36
Launching in a Zmap 37
Zmap keyboard and mouse shortcuts. 38
Tips for a speedier Zmap 40
Blixem 42
Getting Started 43
Running Blixem 43
Input files 43
GFF file 43
Configuration file 45
The Blixem Window 46

Active Strand 47
Big Picture 48
Detail View 49
The toolbar 52
The main menu 53
Hiding sections of the window 54
Operation 56
Navigation 56
Zooming 56
Selections 57
Sorting alignments 59
Fetching sequences 60
Grouping sequences 60
Running dotter 63
Settings 65
Features 65
Display options 65
General settings 66
Columns 66
Grid properties 66
Appearance 67
Key 68
Keyboard shortcuts 69
Dotter 71
Getting Started 72
Running Dotter 72
The Dotter Windows 75
The dot-plot window 75
The alignment tool 76
Greyramp tool 77
Main menu 78
File menu 78
Edit menu 78
View menu 78
Help menu 80
Settings 81
Keyboard shortcuts 83
Annotation resources 84

Blank

Otterlace User Manual

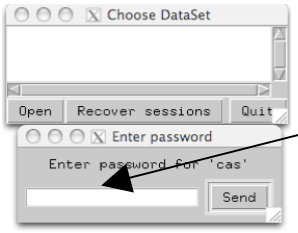
Written by Charles Steward and Laurens WIlming
(cas@sanger.ac.uk, lw2@sanger.ac.uk)
Wellcome Trust Sanger Institute

Otterlace

Otterlace is an interactive, graphical client, which uses a local acedb database with Zmap and perl/Tk tools to curate genomic annotation. Annotation is stored in an extended Ensembl schema (the "otter" database), which presents the annotator with contiguous regions of a chromosome. The acedb database provides local persistent storage, so that if the software or desktop machine crashes, reboots or is exited, the editing session can be recovered. Since all communication goes through the Sanger web server, annotators can work wherever there is a network connection.

Starting an Otterlace Session

Type: **otterlace &** in a terminal window. If you are using Mac OS X, double-click on the otterlace icon. You will be required to authorise your session by entering your password. If you experience any problems, email anacode@sanger.ac.uk

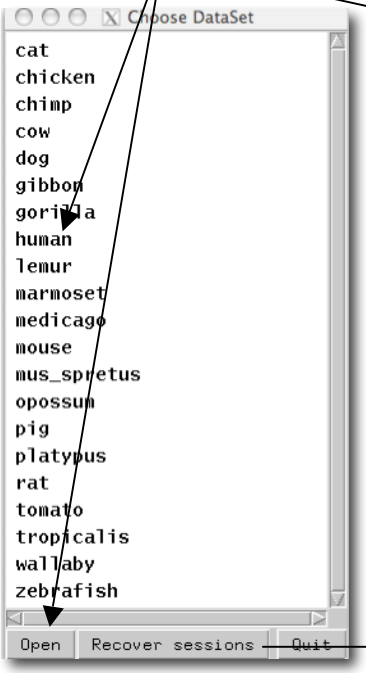


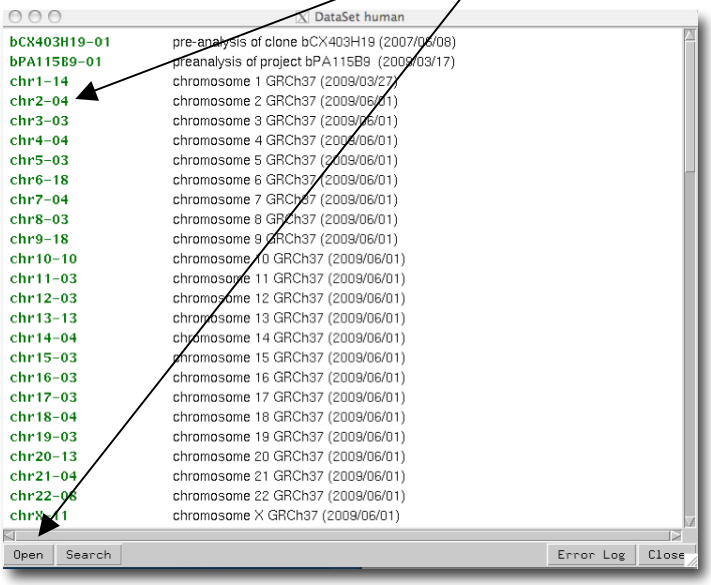
1) Enter your password in the box and click on **Send**.

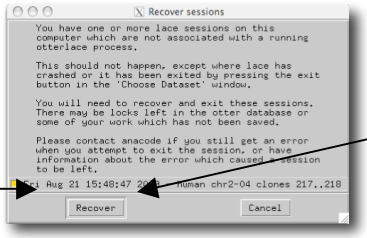
DataSet chooser

3) Select the dataset using left click. Then click on **Open**.

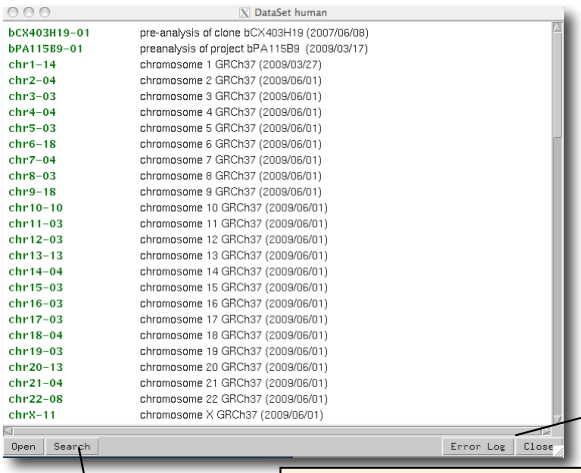
2) Select the species using left click and click on Open or just double click. This will open the **DataSet** window.



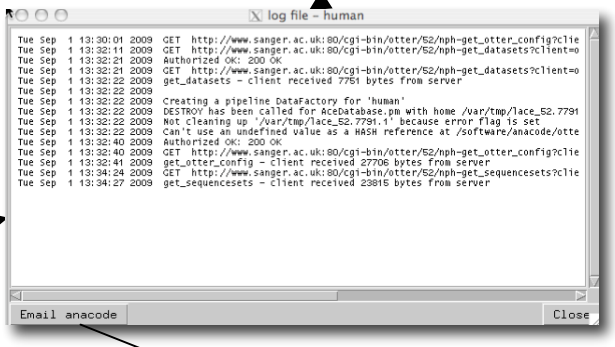

DataSet human
bCX403H19-01 pre-analysis of clone bCX403H19 (2007/06/08)
bPA115B9-01 preanalysis of project bPA115B9 (2008/03/17)
chr1-14 chromosome 1 GRCh37 (2009/03/27)
chr2-04 chromosome 2 GRCh37 (2009/06/01)
chr3-03 chromosome 3 GRCh37 (2009/06/01)
chr4-04 chromosome 4 GRCh37 (2009/06/01)
chr5-03 chromosome 5 GRCh37 (2009/06/01)
chr6-18 chromosome 6 GRCh37 (2009/06/01)
chr7-04 chromosome 7 GRCh37 (2009/06/01)
chr8-03 chromosome 8 GRCh37 (2009/06/01)
chr9-18 chromosome 9 GRCh37 (2009/06/01)
chr10-10 chromosome 10 GRCh37 (2009/06/01)
chr11-03 chromosome 11 GRCh37 (2009/06/01)
chr12-03 chromosome 12 GRCh37 (2009/06/01)
chr13-13 chromosome 13 GRCh37 (2009/06/01)
chr14-04 chromosome 14 GRCh37 (2009/06/01)
chr15-03 chromosome 15 GRCh37 (2009/06/01)
chr16-03 chromosome 16 GRCh37 (2009/06/01)
chr17-03 chromosome 17 GRCh37 (2009/06/01)
chr18-04 chromosome 18 GRCh37 (2009/06/01)
chr19-03 chromosome 19 GRCh37 (2009/06/01)
chr20-13 chromosome 20 GRCh37 (2009/06/01)
chr21-04 chromosome 21 GRCh37 (2009/06/01)
chr22-08 chromosome 22 GRCh37 (2009/06/01)
chrX-11 chromosome X GRCh37 (2009/06/01)
Open
Search
Error Log
Close


Recover sessions
You have one or more lace sessions on this computer which are not associated with a running otterlace process.
This should not happen, except where lace has crashed or it has been exited by pressing the exit button in the 'Choose Dataset' window.
You will need to recover and exit these sessions. There may be locks left in the otter database or some of your work which has not been saved.
Please contact anacode if you still get an error when you attempt to exit the session, or have information about the error which caused a session to be left.
Recover
Cancel

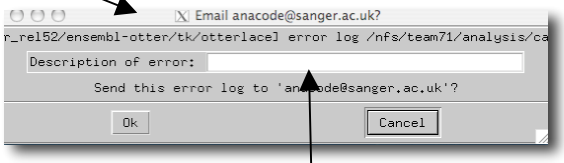
This allows you to recover sessions that have crashed, or when lace has been exited by pressing the **Quit** button in the **Choose Dataset** window. This window will appear automatically when opening a new otterlace session and previous sessions are still present.

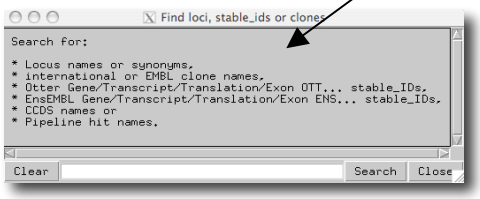

DataSet human
bCX403H19-01 pre-analysis of clone bCX403H19 (2007/06/08)
bPA115B9-01 preanalysis of project bPA115B9 (2009/03/17)
chr1-14 chromosome 1 GRCh37 (2009/03/27)
chr2-04 chromosome 2 GRCh37 (2009/06/01)
chr3-03 chromosome 3 GRCh37 (2009/06/01)
chr4-04 chromosome 4 GRCh37 (2009/06/01)
chr5-03 chromosome 5 GRCh37 (2009/06/01)
chr6-18 chromosome 6 GRCh37 (2009/06/01)
chr7-04 chromosome 7 GRCh37 (2009/06/01)
chr8-03 chromosome 8 GRCh37 (2009/06/01)
chr9-18 chromosome 9 GRCh37 (2009/06/01)
chr10-10 chromosome 10 GRCh37 (2009/06/01)
chr11-03 chromosome 11 GRCh37 (2009/06/01)
chr12-03 chromosome 12 GRCh37 (2009/06/01)
chr13-13 chromosome 13 GRCh37 (2009/06/01)
chr14-04 chromosome 14 GRCh37 (2009/06/01)
chr15-03 chromosome 15 GRCh37 (2009/06/01)
chr16-03 chromosome 16 GRCh37 (2009/06/01)
chr17-03 chromosome 17 GRCh37 (2009/06/01)
chr18-04 chromosome 18 GRCh37 (2009/06/01)
chr19-03 chromosome 19 GRCh37 (2009/06/01)
chr20-13 chromosome 20 GRCh37 (2009/06/01)
chr21-04 chromosome 21 GRCh37 (2009/06/01)
chr22-08 chromosome 22 GRCh37 (2009/06/01)
chrX-11 chromosome X GRCh37 (2009/06/01)
Open
Search
Error Log
Close

Otterlace software and/or database problems are shown in the **Error Log**.

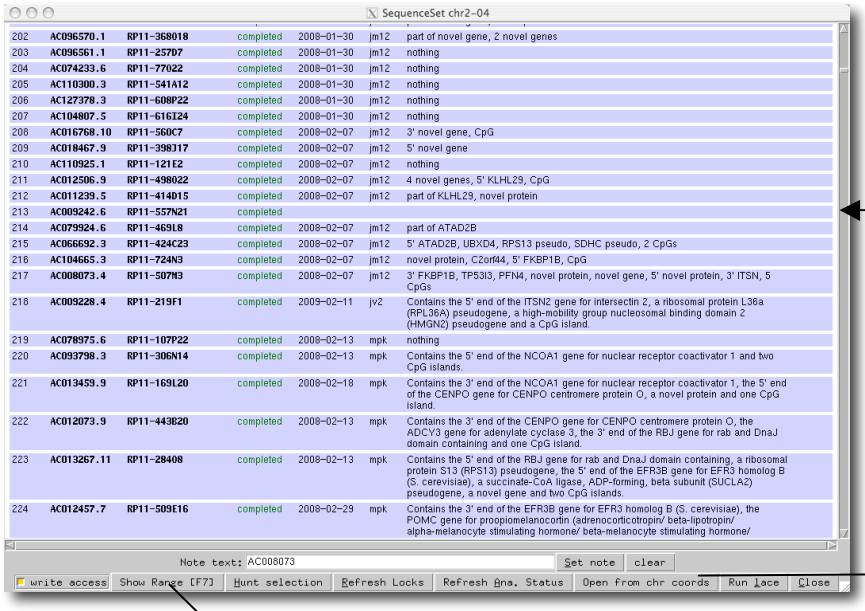

log file - human
Tue Sep 1 13:30:01 2009 GET http://www.sanger.ac.uk:80/cgi-bin/otter/52/nph-get_otter_config?clie
Tue Sep 1 13:32:11 2009 GET http://www.sanger.ac.uk:80/cgi-bin/otter/52/nph-get_datasets?client=o
Tue Sep 1 13:32:21 2009 Authorized OK: 200 OK
Tue Sep 1 13:32:21 2009 GET http://www.sanger.ac.uk:80/cgi-bin/otter/52/nph-get_datasets?client=o
Tue Sep 1 13:32:22 2009 get_datasets - client received 7751 bytes from server
Tue Sep 1 13:32:22 2009
Tue Sep 1 13:32:22 2009 Creating a pipeline DataFactory for 'human'
Tue Sep 1 13:32:22 2009 DESTROY has been called for AceDatabase.pm with home /var/tmp/lace_52.7791
Tue Sep 1 13:32:22 2009 Not cleaning up '/var/tmp/lace_52.7791.1' because error flag is set
Tue Sep 1 13:32:22 2009 Can't use an undefined value as a HASH reference at /software/anacode/otte
Tue Sep 1 13:32:40 2009 Authorized OK: 200 OK
Tue Sep 1 13:32:40 2009 GET http://www.sanger.ac.uk:80/cgi-bin/otter/52/nph-get_otter_config?clie
Tue Sep 1 13:32:41 2009 get_otter_config - client received 27706 bytes from server
Tue Sep 1 13:34:24 2009 GET http://www.sanger.ac.uk:80/cgi-bin/otter/52/nph-get_sequencesets?clie
Tue Sep 1 13:34:27 2009 get_sequencesets - client received 23815 bytes from server
Email anacode
Close

The **Search** feature allows you to search the Dataset for any feature such as Otter ID, gene name etc.

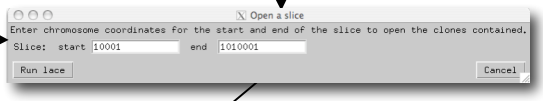


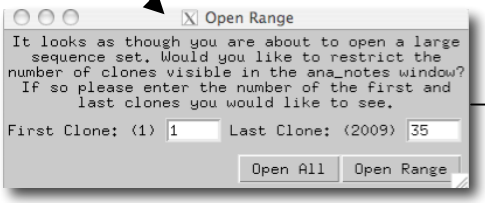


An option to email anacode with the errors is provided to facilitate a diagnosis. Always include a "useful" description in the email!


SequenceSet chr2-04
202 AC096570.1 RP11-368I18 completed 2008-01-30 jm12 part of novel gene, 2 novel genes
203 AC096561.1 RP11-257D7 completed 2008-01-30 jm12 nothing
204 AC074233.6 RP11-77O22 completed 2008-01-30 jm12 nothing
205 AC110300.3 RP11-541A12 completed 2008-01-30 jm12 nothing
206 AC127378.3 RP11-608P22 completed 2008-01-30 jm12 nothing
207 AC104807.5 RP11-616I24 completed 2008-01-30 jm12 nothing
208 AC016768.10 RP11-560C7 completed 2008-02-07 jm12 3' novel gene, CpG
209 AC018467.9 RP11-398I17 completed 2008-02-07 jm12 5' novel gene
210 AC110925.1 RP11-121E2 completed 2008-02-07 jm12 nothing
211 AC012506.9 RP11-498O22 completed 2008-02-07 jm12 4 novel genes, 5' KLHL29, CpG
212 AC011239.5 RP11-414D15 completed 2008-02-07 jm12 part of KLHL29, novel protein
213 AC009242.6 RP11-557N21 completed
214 AC079924.6 RP11-469I8 completed 2008-02-07 jm12 part of ATAD2B
215 AC066692.3 RP11-424C23 completed 2008-02-07 jm12 5' ATAD2B, UBXD4, RPS13 pseudo, SDHC pseudo, 2 CpGs
216 AC104665.3 RP11-724N3 completed 2008-02-07 jm12 novel protein, C2orf44, 5' FKBP1B, CpG
217 AC008073.4 RP11-507M3 completed 2008-02-07 jm12 3' FKBP1B, TP53I3, PFN4, novel protein, novel gene, 5' novel protein, 3' ITSN, 5 CpGs
218 AC009228.4 RP11-219F1 completed 2009-02-11 jv2 Contains the 5' end of the ITSN2 gene for intersectin 2, a ribosomal protein L36a (RPL36A) pseudogene, a high-mobility group nucleosomal binding domain 2 (HMGN2) pseudogene and a CpG island.
219 AC078975.6 RP11-107P22 completed 2008-02-13 mpk nothing
220 AC093798.3 RP11-306N14 completed 2008-02-13 mpk Contains the 5' end of the NCOA1 gene for nuclear receptor coactivator 1 and two CpG islands.
221 AC013459.9 RP11-169L20 completed 2008-02-18 mpk Contains the 3' end of the NCOA1 gene for nuclear receptor coactivator 1, the 5' end of the CENPO gene for CENPO centromere protein O, a novel protein and one CpG island.
222 AC012073.9 RP11-443B20 completed 2008-02-13 mpk Contains the 3' end of the CENPO gene for CENPO centromere protein O, the ADCY3 gene for adenylate cyclase 3, the 3' end of the RBJ gene for rab and DnaJ domain containing and one CpG island.
223 AC013267.11 RP11-28A8 completed 2008-02-13 mpk Contains the 5' end of the RBJ gene for rab and DnaJ domain containing, a ribosomal protein S13 (RPS13) pseudogene, the 5' end of the EFR3B gene for EFR3 homolog B (S. cerevisiae), a succinate-CoA ligase, ADP-forming, beta subunit (SUCLA2) pseudogene, a novel gene and two CpG islands.
224 AC012457.7 RP11-509E16 completed 2008-02-29 mpk Contains the 3' end of the EFR3B gene for EFR3 homolog B (S. cerevisiae), the POMC gene for proopiomelanocortin (adrenocorticotropin/ beta-lipotropin/ alpha-melanocyte stimulating hormone/ beta-melanocyte stimulating hormone/
Note text: AC008073
Set note
clear
write access
Show Range [F7]
Hunt selection
Refresh Locks
Refresh Ana. Status
Open from chr coords
Run lace
Close

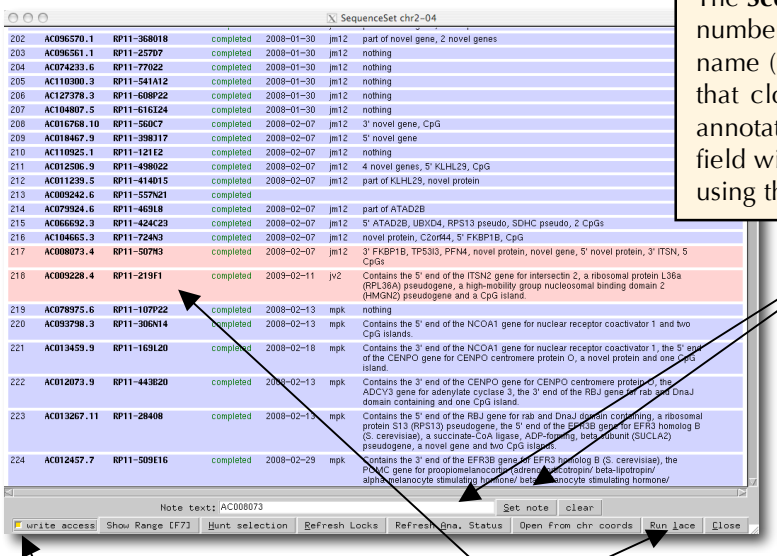
4) The **SequenceSet** window appears, (also known as **Ana_notes)**. It shows remarks that can be added using the entry field at the bottom to help track annotation progress. This window also allows you to either open the whole contig range in one scrollable window or open a selected range of your choice.

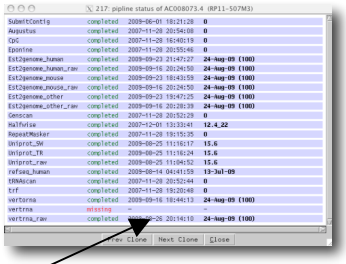

Open a slice
Enter chromosome coordinates for the start and end of the slice to open the clones contained.
Slice: start 10001 end 1010001
Run lace
Cancel


Open Range
It looks as though you are about to open a large sequence set. Would you like to restrict the number of clones visible in the ana_notes window? If so please enter the number of the first and last clones you would like to see.
First Clone: (1) 1
Last Clone: (2009) 35
Open All
Open Range

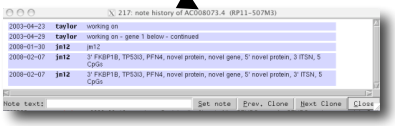
These options allow you to open specific regions and are designed to make opening clones quicker.

The **SequenceSet** window columns show: clone number, accession number, internal project name (where appropriate), pipeline status, date that clone entry in this window was updated, annotator responsible for update and free text field with notes about the clone content entered using the **Note text** box.





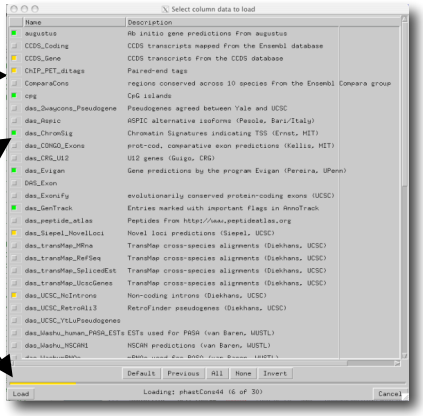
Right click on clone to show report of pipeline status. Double clicking on a clone shows history of edits.



write access can be turned off by clicking on the yellow button.

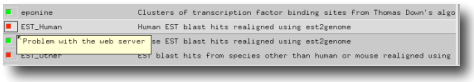
5) A clone can be selected using the left mouse button. Use the shift button to select multiple clones. Selected clones become a nice salmon colour. Now click **Run lace**.

6) The **Select column data to load** box appears next, which allows you to select the analysis and features you wish to see in Zmap. Fewer selected columns will mean a shorter period of time required to open your clones. Selected columns have a yellow box next to them.



8) A yellow progress bar shows the status of data loading. Yellow boxes turn **green** when columns are loaded successfully.

Failed columns turn **red** (mouse over for details). Click button to return status to yellow to retry.

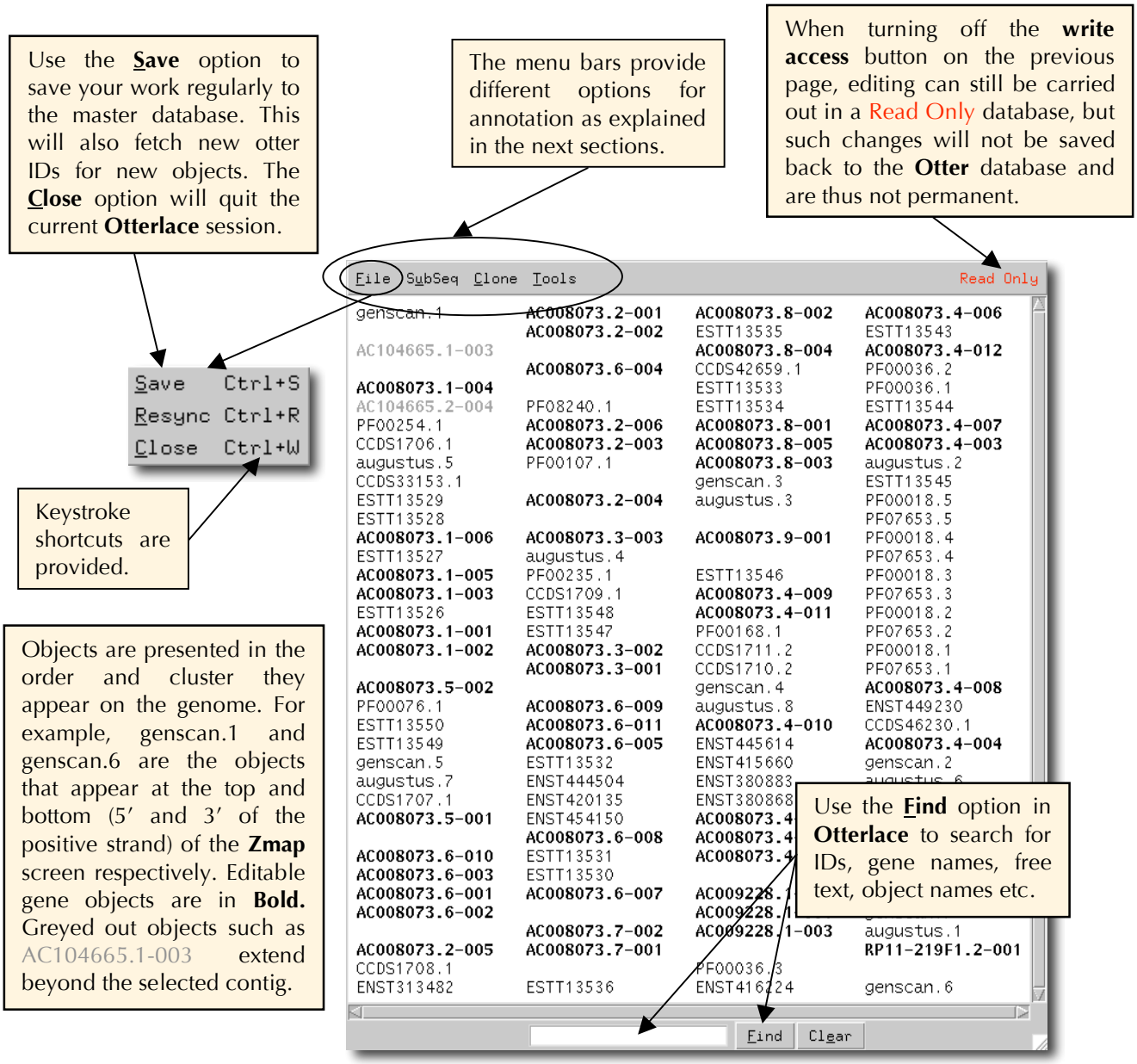


7) Click **load** to run **Otterlace** and **Zmap.**

Further description information for the **column data** can be found here:
http://scratchy.internal.sanger.ac.uk/wiki/index.php/Otterlace_filter_descriptions

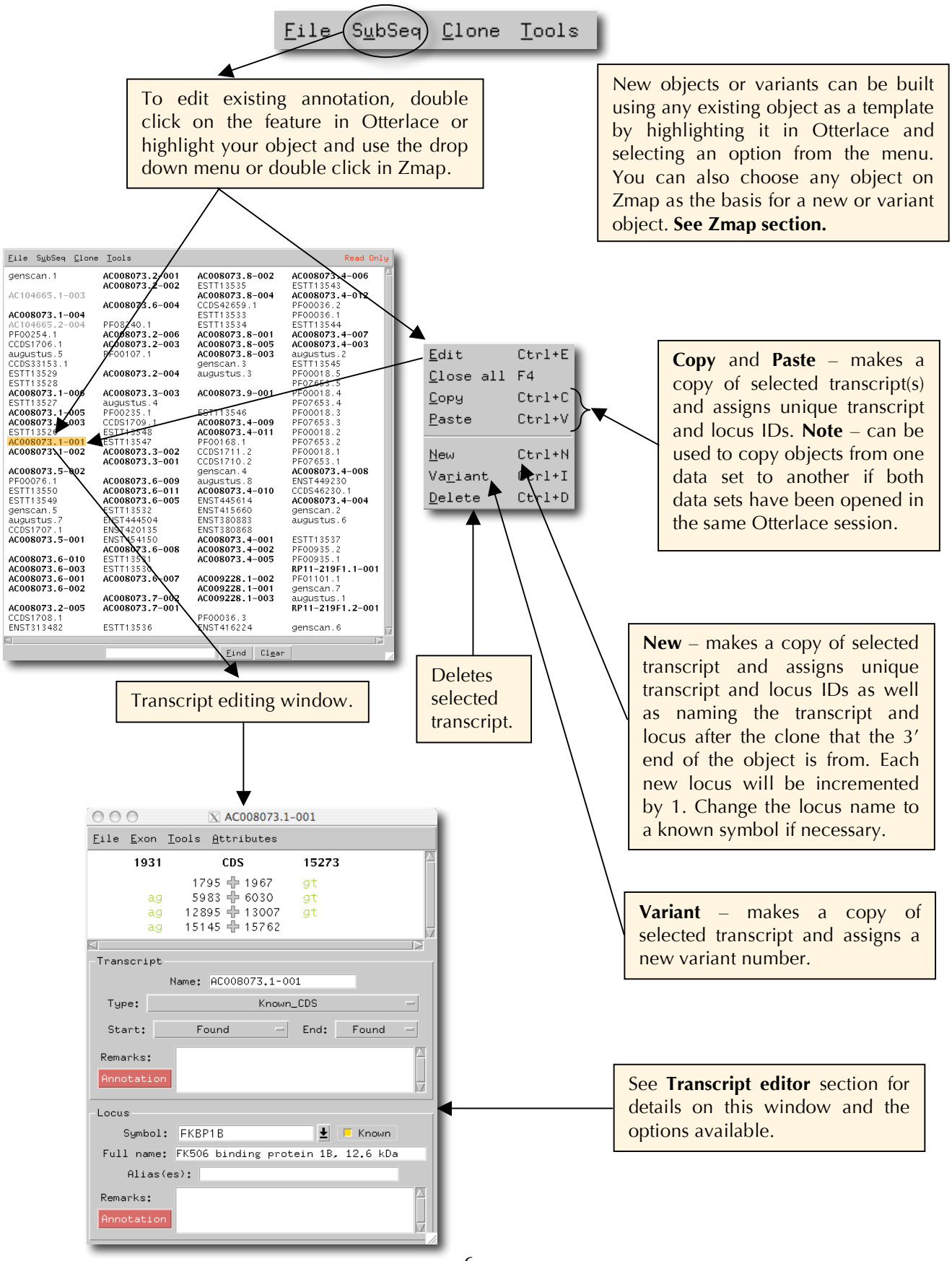
Transcript chooser section

File menu: Manage the Otterlace editing session.

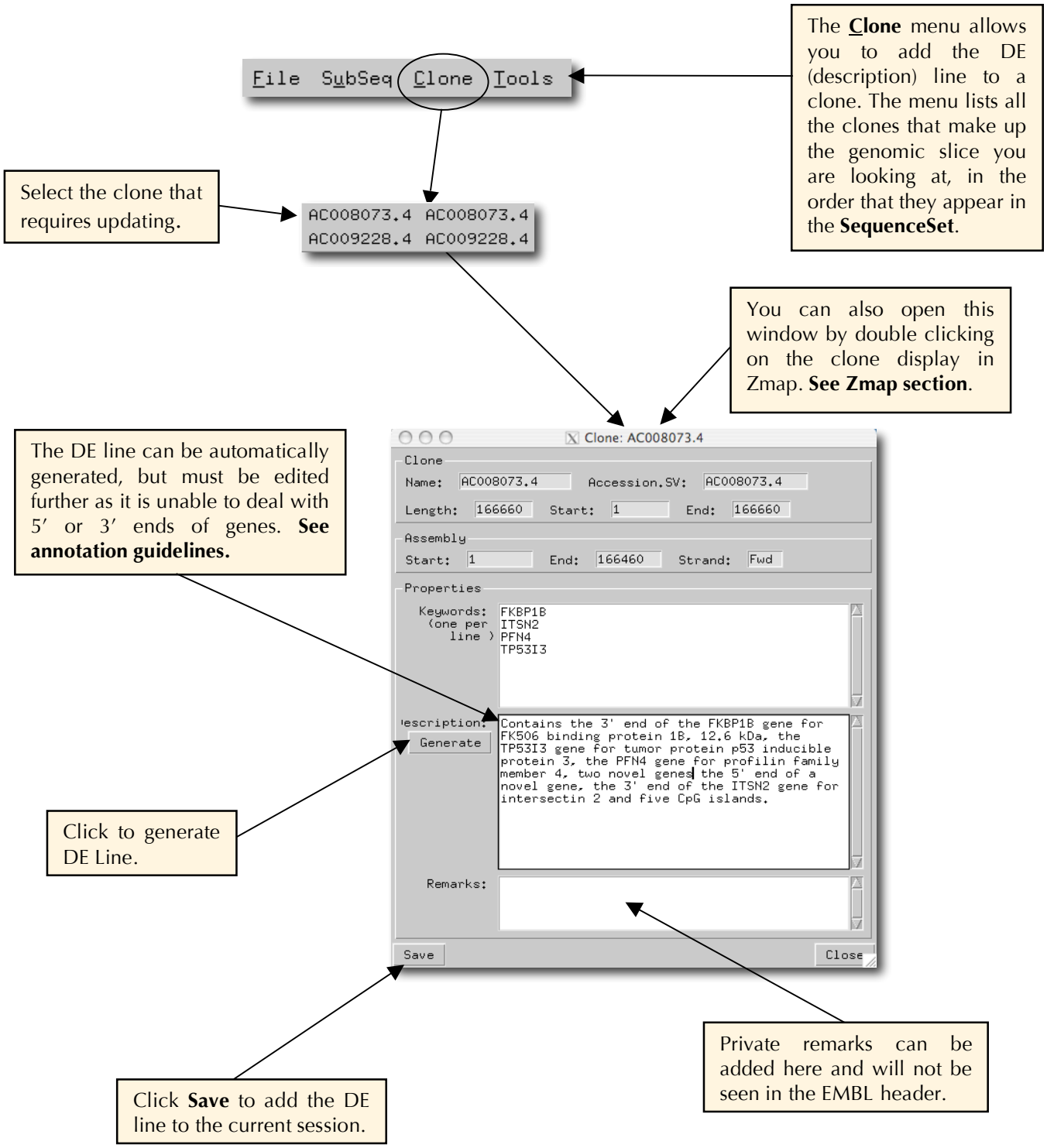

Use the **Save** option to save your work regularly to the master database. This will also fetch new otter IDs for new objects. The **Close** option will quit the current **Otterlace** session.
The menu bars provide different options for annotation as explained in the next sections.
When turning off the **write access** button on the previous page, editing can still be carried out in a Read Only database, but such changes will not be saved back to the **Otter** database and are thus not permanent.
File SubSeq Clone Tools
Read Only
Save Ctrl+S
Resync Ctrl+R
Close Ctrl+W
Keystroke shortcuts are provided.
Objects are presented in the order and cluster they appear on the genome. For example, genscan.1 and genscan.6 are the objects that appear at the top and bottom (5′ and 3′ of the positive strand) of the **Zmap** screen respectively. Editable gene objects are in **Bold.** Greyed out objects such as AC104665.1-003 extend beyond the selected contig.
Use the **Find** option in **Otterlace** to search for IDs, gene names, free text, object names etc.
Find
Clear

The locus and its associated transcripts and exons are attributed stable, versioned database IDs (e.g. OTTHUMG00000017411), generated and tracked within the **Otter** database. Whenever a gene locus is edited the version number will increase and the date of the change will be saved, allowing the user to find out when the annotation was last updated. It should be noted that versioning occurs within the database and such changes are not externally visible. Clearly, it is vital that current Otter IDs are not deleted, only modified, unless the object is no longer valid.

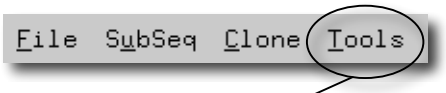
Subseq menu: Editing operations on the transcripts listed in the window.

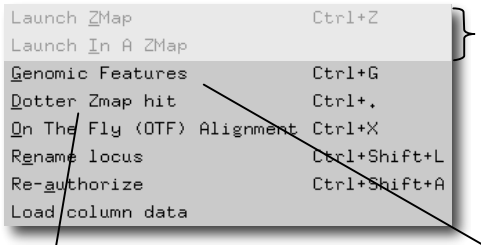

File SubSeq Clone Tools
To edit existing annotation, double click on the feature in Otterlace or highlight your object and use the drop down menu or double click in Zmap.
New objects or variants can be built using any existing object as a template by highlighting it in Otterlace and selecting an option from the menu. You can also choose any object on Zmap as the basis for a new or variant object. See Zmap section.
Edit Ctrl+E
Close all F4
Copy Ctrl+C
Paste Ctrl+V
New Ctrl+N
Variant Ctrl+I
Delete Ctrl+D
Copy and Paste – makes a copy of selected transcript(s) and assigns unique transcript and locus IDs. Note – can be used to copy objects from one data set to another if both data sets have been opened in the same Otterlace session.
Deletes selected transcript.
New – makes a copy of selected transcript and assigns unique transcript and locus IDs as well as naming the transcript and locus after the clone that the 3′ end of the object is from. Each new locus will be incremented by 1. Change the locus name to a known symbol if necessary.
Variant – makes a copy of selected transcript and assigns a new variant number.
Transcript editing window.
AC008073.1-001
File Exon Tools Attributes
1931 CDS 15273
1795 1967
5983 6030
12895 13007
15145 15762
Transcript
Name: AC008073.1-001
Type: Known_CDS
Start: Found
End: Found
Remarks:
Annotation
Locus
Symbol: FKBP1B
Known
Full name: FK506 binding protein 1B, 12.6 kDa
Alias(es):
Remarks:
Annotation
See Transcript editor section for details on this window and the options available.

Clone menu: Edit properties of each of the clones (one or many) opened in the otterlace editing session.

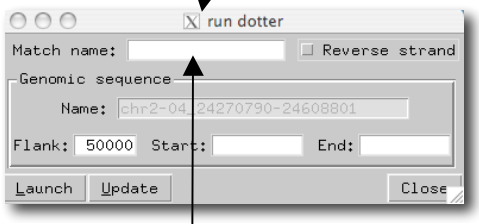

File SubSeq Clone Tools
The Clone menu allows you to add the DE (description) line to a clone. The menu lists all the clones that make up the genomic slice you are looking at, in the order that they appear in the SequenceSet.
Select the clone that requires updating.
AC008073.4 AC008073.4
AC009228.4 AC009228.4
You can also open this window by double clicking on the clone display in Zmap. See Zmap section.
The DE line can be automatically generated, but must be edited further as it is unable to deal with 5′ or 3′ ends of genes. See annotation guidelines.
Clone: AC008073.4
Clone
Name: AC008073.4
Accession.SV: AC008073.4
Length: 166660
Start: 1
End: 166660
Assembly
Start: 1
End: 166460
Strand: Fwd
Properties
Keywords: (one per line)
FKBP1B
ITSN2
PFN4
TP53I3
Description:
Generate
Contains the 3' end of the FKBP1B gene for FK506 binding protein 1B, 12.6 kDa, the TP53I3 gene for tumor protein p53 inducible protein 3, the PFN4 gene for profilin family member 4, two novel genes the 5' end of a novel gene, the 3' end of the ITSN2 gene for intersectin 2 and five CpG islands.
Remarks:
Save
Close
Click to generate DE Line.
Click Save to add the DE line to the current session.
Private remarks can be added here and will not be seen in the EMBL header.

Tools menu: Useful things to run on the genomic sequence being annotated.

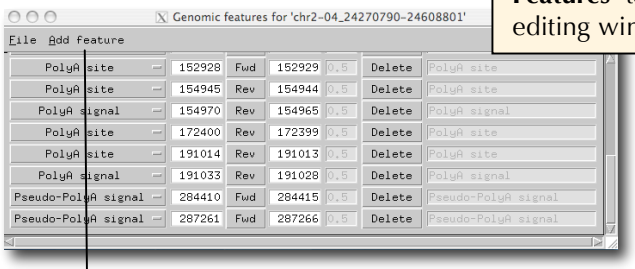




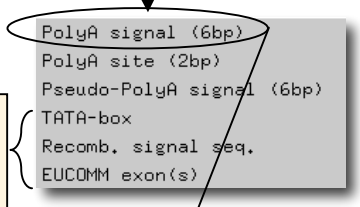
Use this to relaunch Zmap if it was accidentally closed. For Zmap options, see **Zmap** section.



Dotter alignment of any selected homology in paste buffer to object. See section on **Dotter.**


Genomic features for 'chr2-04_24270790-24608801'
File Add feature
PolyA site 152928 Fwd 152929 0.5 Delete PolyA site
PolyA site 154945 Rev 154944 0.5 Delete PolyA site
PolyA signal 154970 Rev 154965 0.5 Delete PolyA signal
PolyA site 172400 Rev 172399 0.5 Delete PolyA site
PolyA site 191014 Rev 191013 0.5 Delete PolyA site
PolyA signal 191033 Rev 191028 0.5 Delete PolyA signal
Pseudo-PolyA signal 284410 Fwd 284415 0.5 Delete Pseudo-PolyA signal
Pseudo-PolyA signal 287261 Fwd 287266 0.5 Delete Pseudo-PolyA signal

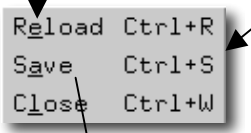
Select **Genomic Features** to bring up editing window.

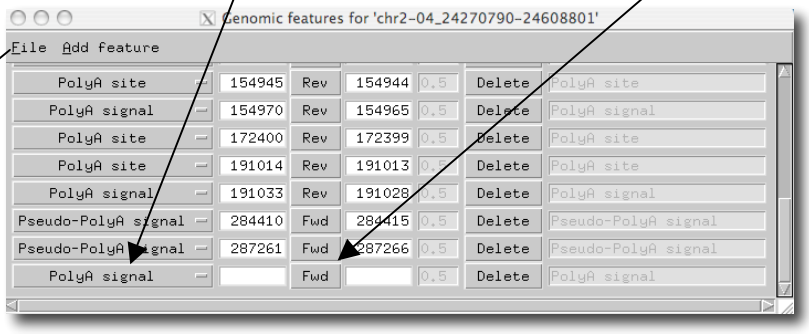


Some features are project specific and will be defined when working on that project.

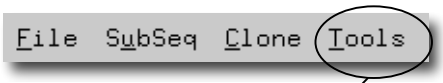
From the **Add feature** menu select the type of feature you want to add. This will then appear in the main box. For polyA features only one of the coordinates needs to be entered as the other is calculated automatically. If necessary, click to toggle the direction (Fwd/Rev). **Select strand before entering coordinates**.

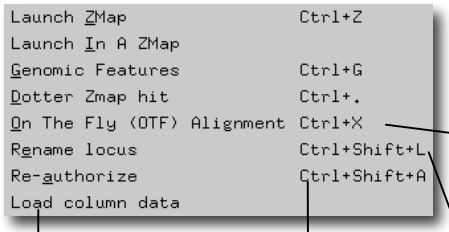
Reload reverts features back to the last save.



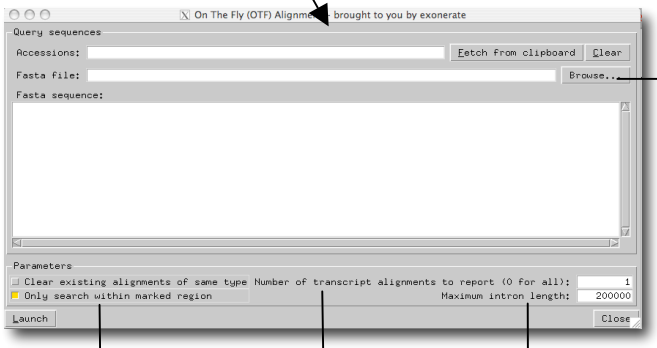

Genomic features for 'chr2-04_24270790-24608801'
File Add feature
PolyA site 154945 Rev 154944 0.5 Delete PolyA site
PolyA signal 154970 Rev 154965 0.5 Delete PolyA signal
PolyA site 172400 Rev 172399 0.5 Delete PolyA site
PolyA site 191014 Rev 191013 0.5 Delete PolyA site
PolyA signal 191033 Rev 191028 0.5 Delete PolyA signal
Pseudo-PolyA signal 284410 Fwd 284415 0.5 Delete Pseudo-PolyA signal
Pseudo-PolyA signal 287261 Fwd 287266 0.5 Delete Pseudo-PolyA signal
PolyA signal Fwd 0.5 Delete PolyA signal

Once the coordinates have been entered, select **Save** from the main window to see the features in Zmap.





On The Fly (OTF) alignment uses exonerate to align sequences to Zmap. These can be single sequences, multiple sequences highlighted in Zmap, missing accession numbers or a fasta file of one or many sequences. Data can be entered in all three of the fields in the OTF window at the same time to search on accession(s), from a file and a seqtext area. Results are dynamically loaded onto Zmap to the right or left of the clone lines **(see later Zmap section)**, depending on orientation.

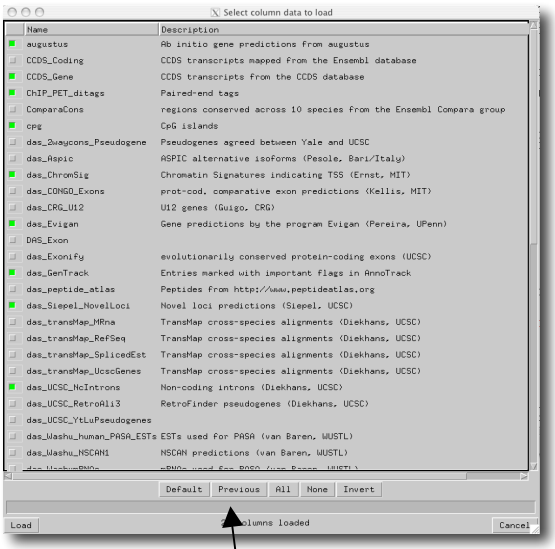


Re-authorize allows you re-establish connection to the database if login expires. This can occur if session has been running for a few days.

Limit search to the marked region in Zmap

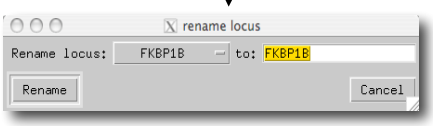
Set this box to "1" to search for the best match or set to "0" to search for all matches.

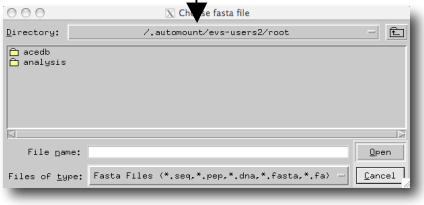
Use this window to increase the window of search for genes with large introns.



Renames a locus to a new locus name.

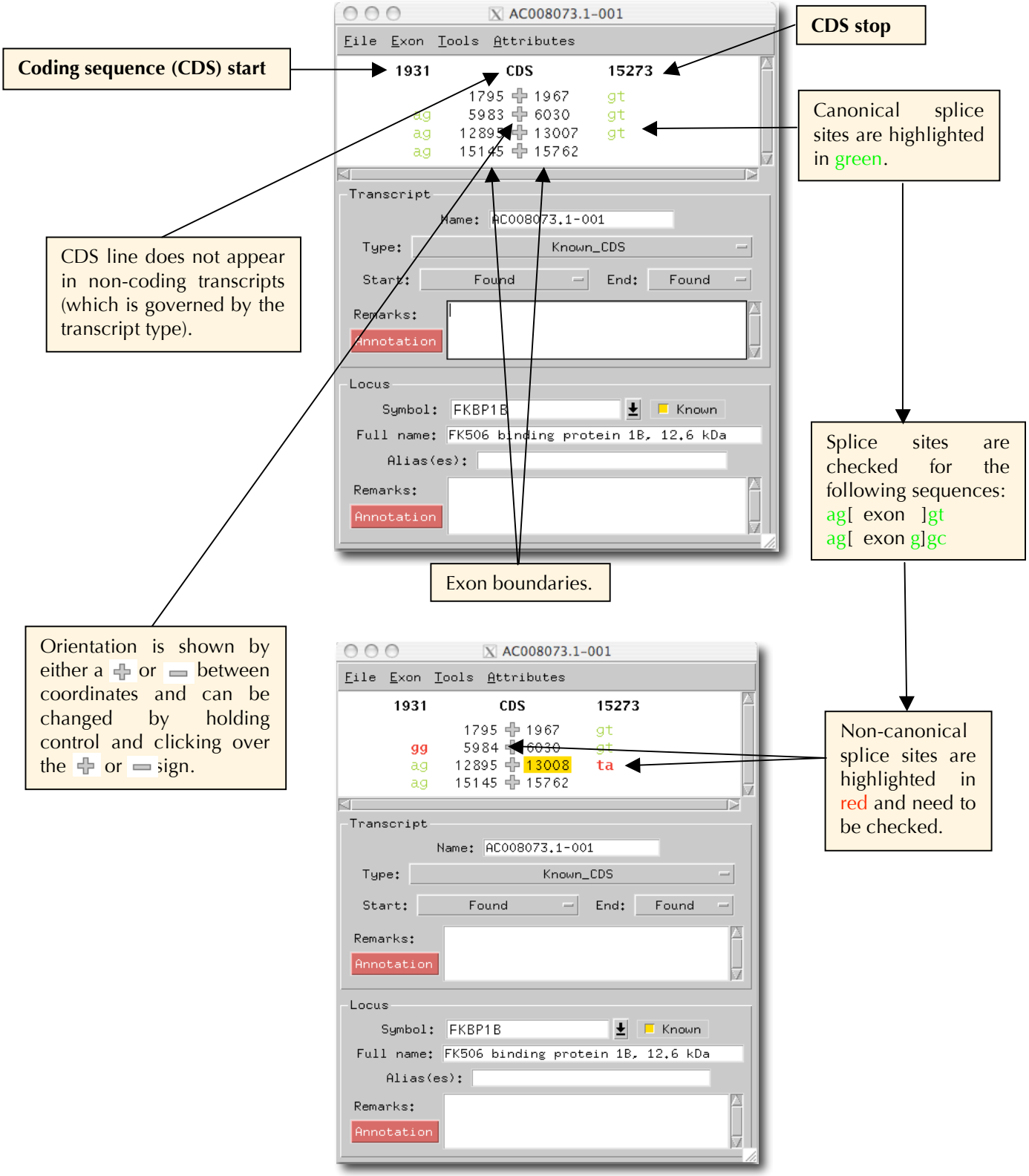
Browse local directory for sequence files.





Load column data gives you the option to load in further column data to a session that is already open.

Transcript editor section



Changing the coordinates can be done a number of ways:
a) Copy coordinates from Blixem (**see section on Blixem**) or select a block of your choice (exon, homology, …) in Zmap and paste coordinates in white space to create new exon(s).
b) Or select existing exon(s) and paste to create copies to edit.
c) Paste over existing coordinate to replace old with new.
d) Or select coordinate and use up and down cursor key to change value.
e) Or select coordinate and delete numbers with backspace key and type in new numbers.
Note: Pasting is done by pressing the middle mouse button (often the scroll wheel).

Transcript type (see **Annotation Guidelines**).

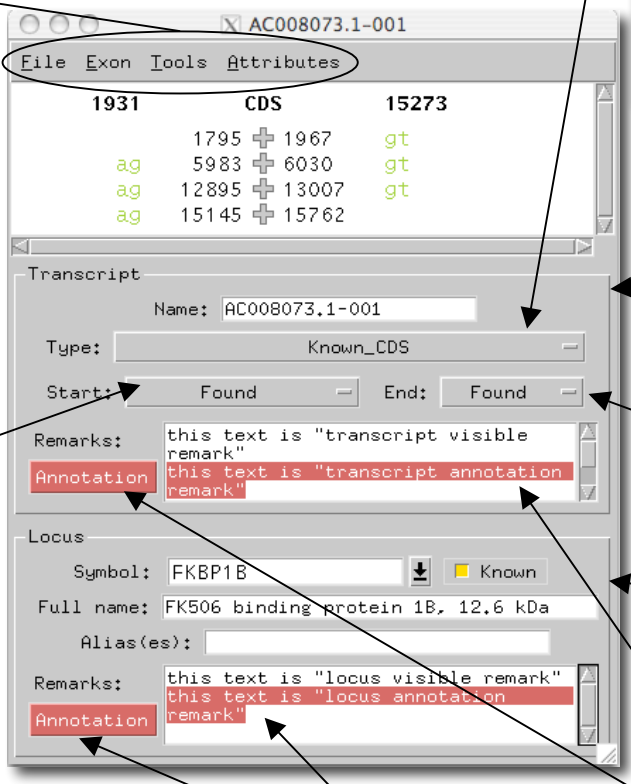
Coding
- Known_CDS
- Novel_CDS
- Putative_CDS
- Nonsense_mediated_decay

Transcript
- Non_coding
- Ambiguous_ORF
- Retained_intron
- Antisense
- Disrupted_domain
- IG_segment
- IG_gene
- Putative

Pseudogene
- Processed_pseudogene
- Unprocessed_pseudogene
- Transcribed_processed_pseudogene
- Transcribed_unprocessed_pseudogene
- Unitary_pseudogene
- Polymorphic_pseudogene
- IG_pseudogene
- Expressed_pseudogene

Transposon
Artifact
TEC
Predicted

The next section describes these menu options.



This section provides information relevant to the **transcript**.

Found
Not Found

Status of translation stop (CDS only).

Found
CDS not found - 1
CDS not found - 2
CDS not found - 3
UTR incomplete

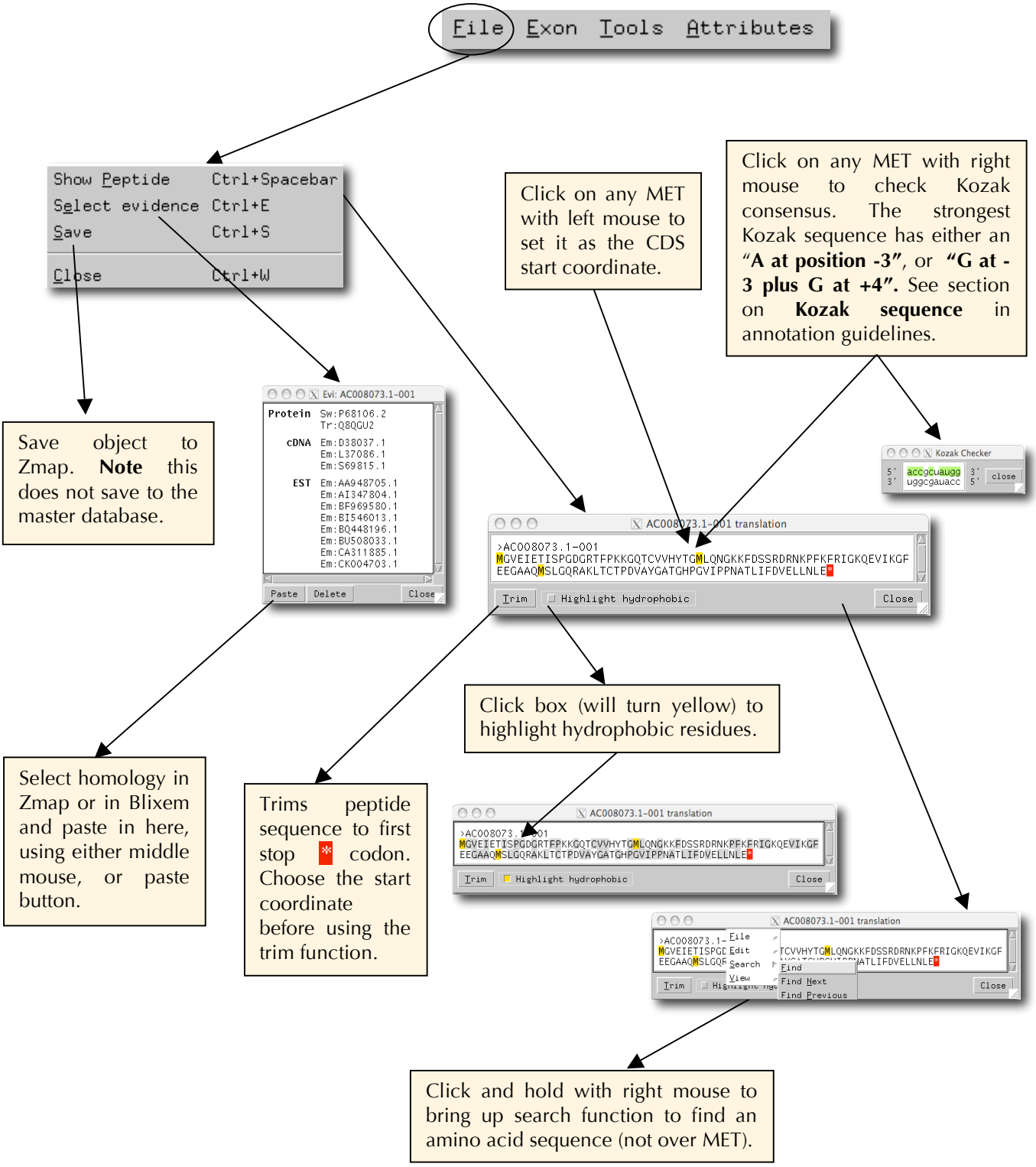
This section provides information relevant to the **locus** (gene).

Status of translation start (CDS only); number indicates translation off-set. The UTR incomplete tag is set if the transcript is cut off within the UTR. For example, if not all of an mRNA used as evidence can be aligned, due to missing genomic sequence etc.

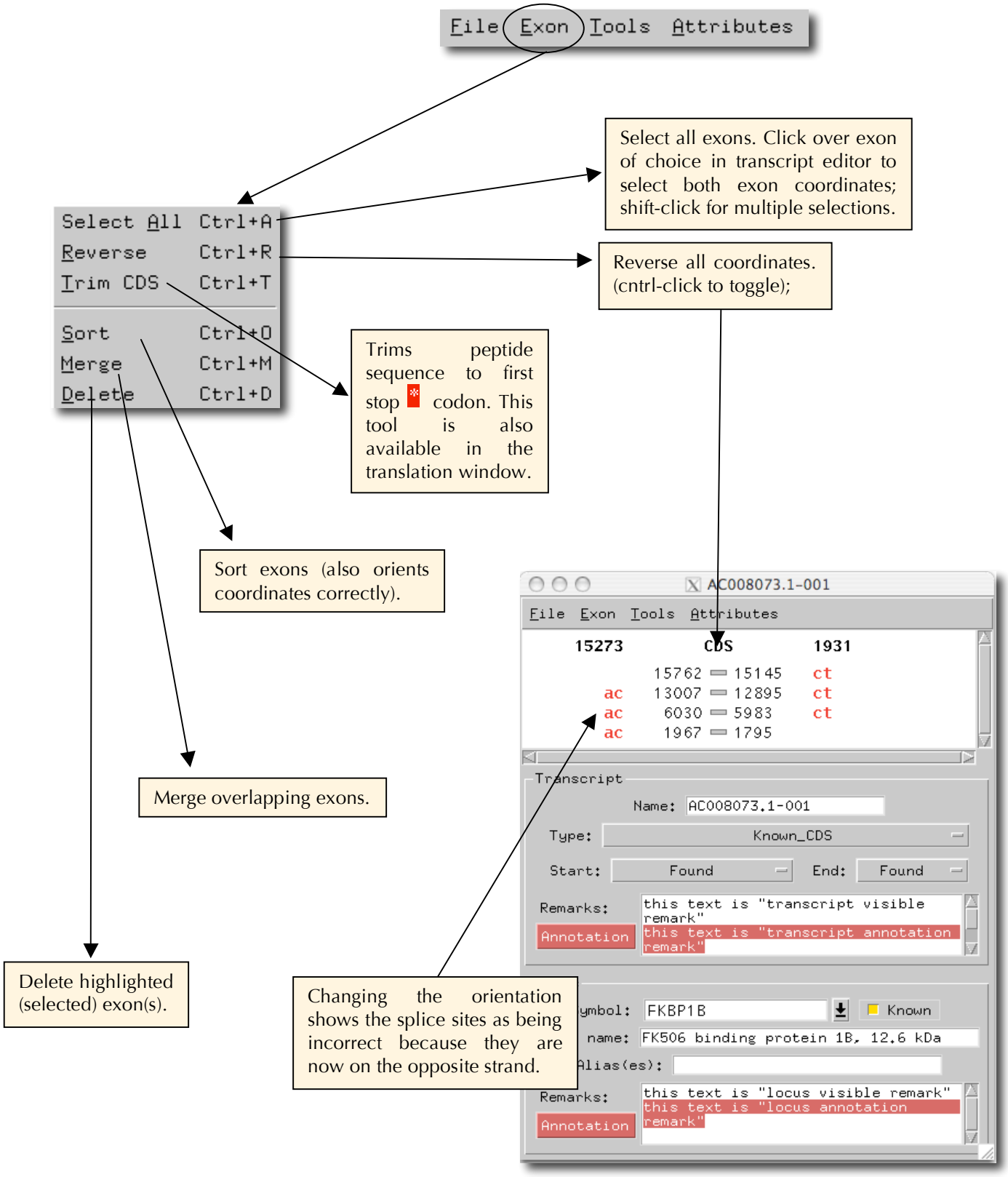
Locus notes. Click on red annotation button to make a comment private, so that it does not appear in the EMBL file.

Transcript notes. Click on red annotation button to make a comment private, so that it does not appear in the EMBL file.

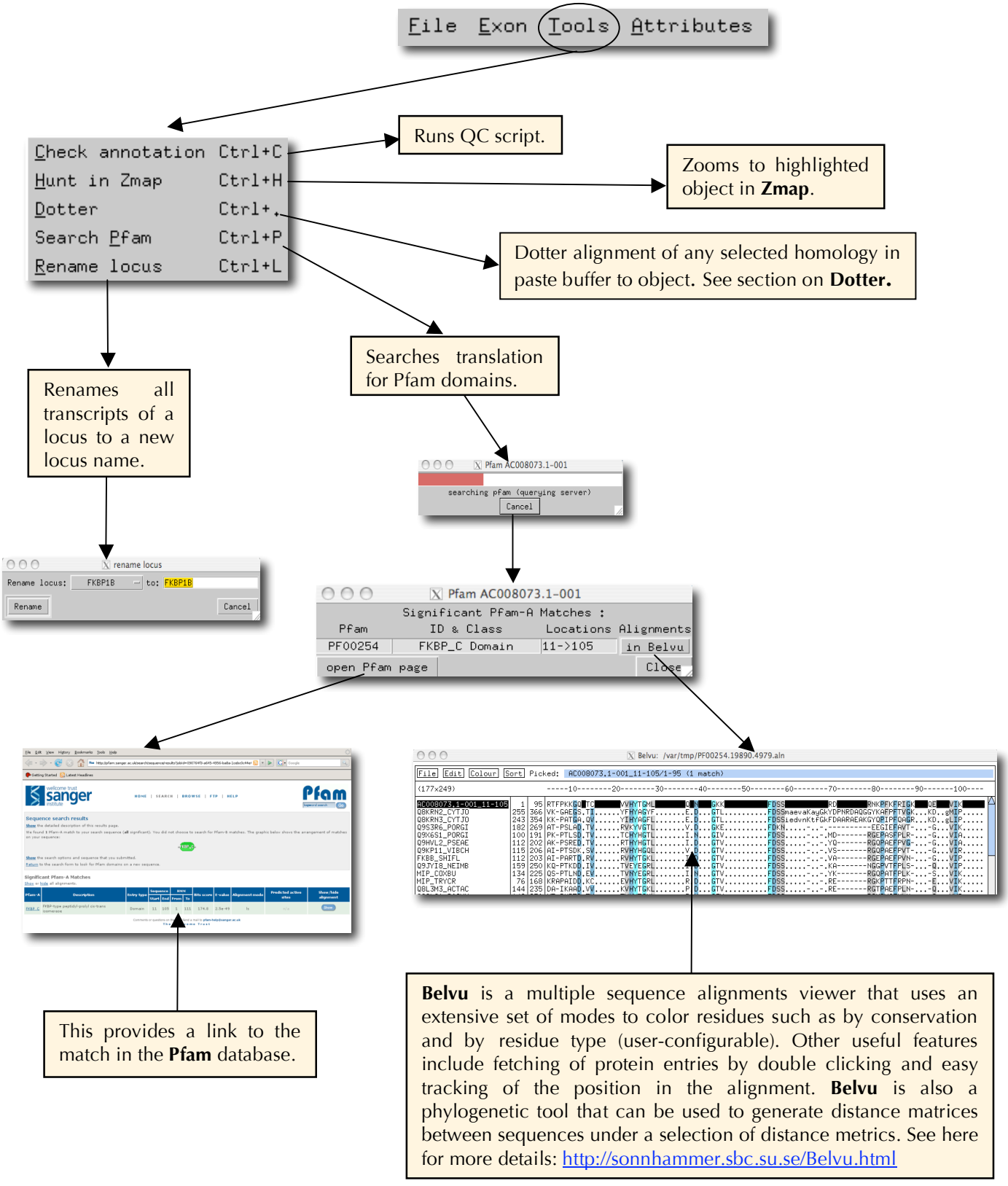
File menu: Saving, closing, plus windows for showing translation and selecting supporting evidence.


File Exon Tools Attributes
Show Peptide Ctrl+Spacebar
Select evidence Ctrl+E
Save Ctrl+S
Close Ctrl+W
Save object to Zmap. **Note** this does not save to the master database.
Evi: AC008073.1-001
Protein Sw:P68106.2 Tr:Q8QGU2
cDNA Em:D38037.1 Em:L37086.1 Em:S69815.1
EST Em:AA948705.1 Em:AI347804.1 Em:BF969580.1 Em:BI546013.1 Em:BQ448196.1 Em:BU508033.1 Em:CA311885.1 Em:CK004703.1
Paste Delete Close
Click on any MET with left mouse to set it as the CDS start coordinate.
Click on any MET with right mouse to check Kozak consensus. The strongest Kozak sequence has either an "**A at position -3"**, or **"G at -3 plus G at +4". See section on **Kozak sequence** in annotation guidelines.
Kozak Checker
close
AC008073.1-001 translation
>AC008073.1-001
Trim Highlight hydrophobic Close
Select homology in Zmap or in Blixem and paste in here, using either middle mouse, or paste button.
Trims peptide sequence to first stop codon. Choose the start coordinate before using the trim function.
Click box (will turn yellow) to highlight hydrophobic residues.
File Edit Search View
Find Find Next Find Previous
Click and hold with right mouse to bring up search function to find an amino acid sequence (not over MET).

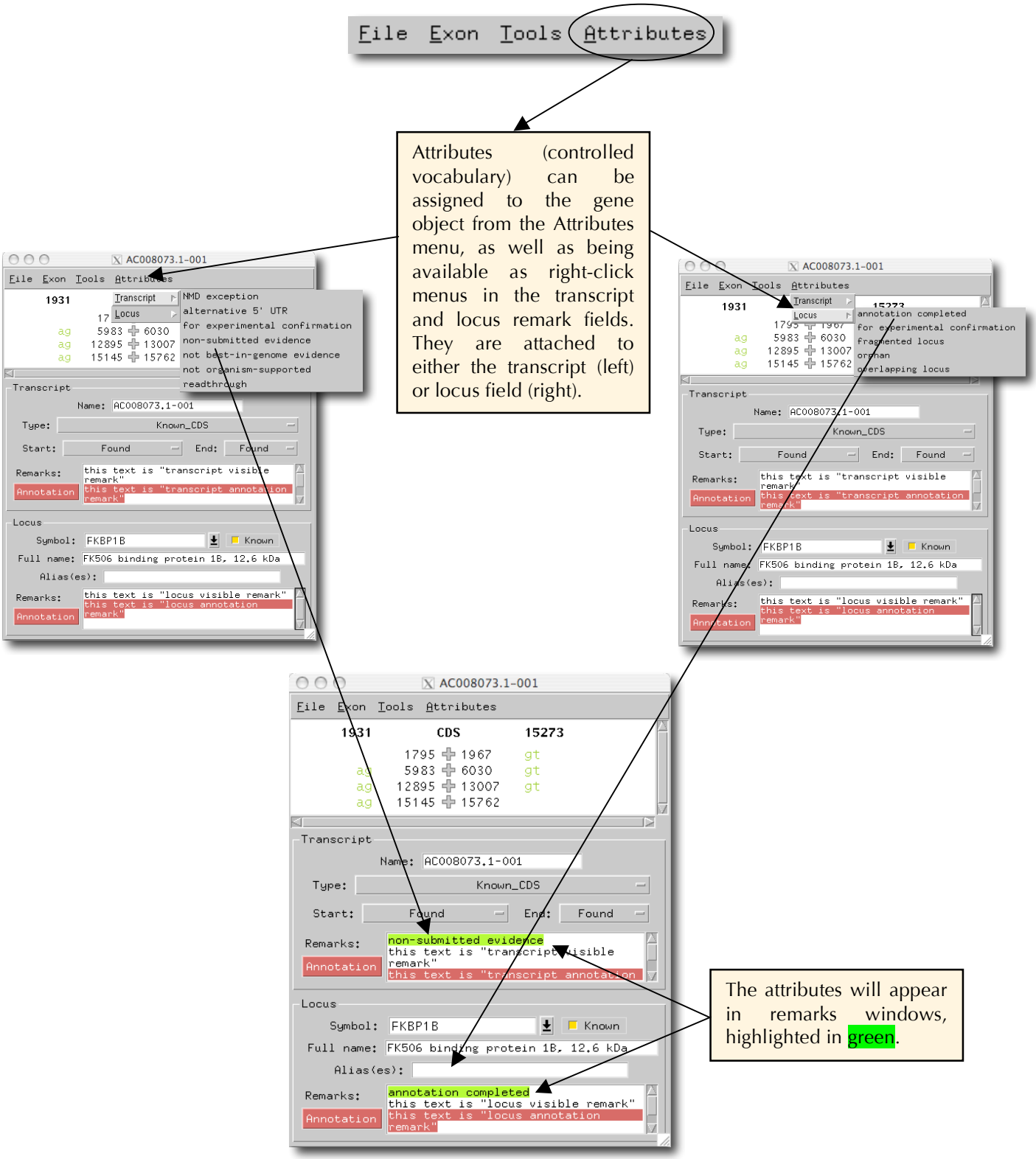
Exon menu: Tools for editing the exons.


File Exon Tools Attributes
Select All Ctrl+A
Reverse Ctrl+R
Trim CDS Ctrl+T
Sort Ctrl+O
Merge Ctrl+M
Delete Ctrl+D
Select all exons. Click over exon of choice in transcript editor to select both exon coordinates; shift-click for multiple selections.
Reverse all coordinates. (cntrl-click to toggle);
Trims peptide sequence to first stop codon. This tool is also available in the translation window.
Sort exons (also orients coordinates correctly).
Merge overlapping exons.
Delete highlighted (selected) exon(s).
Changing the orientation shows the splice sites as being incorrect because they are now on the opposite strand.
AC008073.1-001
File Exon Tools Attributes
15273 CDS 1931
15762 15145 ct
ac 13007 12895 ct
ac 6030 5983 ct
ac 1967 1795
Transcript
Name: AC008073.1-001
Type: Known_CDS
Start: Found End: Found
Remarks:
this text is "transcript visible remark"
Annotation
this text is "transcript annotation remark"
ymbol: FKBP1B
Known
name: FK506 binding protein 1B, 12.6 kDa
Alias(es):
Remarks:
this text is "locus visible remark"
this text is "locus annotation remark"
Annotation

Tools menu: Informative operations to run on the transcript.

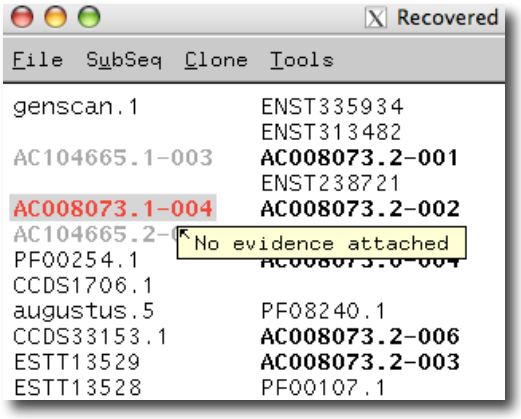

File Exon Tools Attributes
Check annotation Ctrl+C
Hunt in Zmap Ctrl+H
Dotter Ctrl+.
Search Pfam Ctrl+P
Rename locus Ctrl+L
Runs QC script.
Zooms to highlighted object in Zmap.
Dotter alignment of any selected homology in paste buffer to object. See section on Dotter.
Searches translation for Pfam domains.
Renames all transcripts of a locus to a new locus name.
Pfam AC008073.1-001
searching pfam (querying server)
Cancel
rename locus
Rename locus: FKBP1B to: FKBP1B
Rename
Cancel
Pfam AC008073.1-001
Significant Pfam-A Matches :
Pfam ID & Class Locations Alignments
PF00254 FKBP_C Domain 11->105 in Belvu
open Pfam page
Close
Belvu: /var/tmp/PF00254.19890.4979.aln
This provides a link to the match in the Pfam database.
Belvu is a multiple sequence alignments viewer that uses an extensive set of modes to color residues such as by conservation and by residue type (user-configurable). Other useful features include fetching of protein entries by double clicking and easy tracking of the position in the alignment. Belvu is also a phylogenetic tool that can be used to generate distance matrices between sequences under a selection of distance metrics. See here for more details: http://sonnhammer.sbc.su.se/Belvu.html

Attributes menu: Controlled annotation vocabulary for transcript and locus.


File Exon Tools Attributes
Attributes (controlled vocabulary) can be assigned to the gene object from the Attributes menu, as well as being available as right-click menus in the transcript and locus remark fields. They are attached to either the transcript (left) or locus field (right).
AC008073.1-001
NMD exception
alternative 5' UTR
for experimental confirmation
non-submitted evidence
not best-in-genome evidence
not organism-supported
readthrough
annotation completed
for experimental confirmation
fragmented locus
orphan
overlapping locus
non-submitted evidence
annotation completed
The attributes will appear in remarks windows, highlighted in green.

Quality control

Otterlace has a built in annotation checking system that checks all manual annotation as it is being created, as well as existing manual annotation, flagging up inconsistent gene objects in red. If you mouse over the offending gene object, you will see a balloon appear explaining any errors that the checking software has found.



This example shows that gene object AC008073.1-004 has no supporting evidence added to it. The gene object will turn black once the checking software finds no inconsistencies.

The complete list of checks carried out is as follows:

1) No internal stop codons exist in coding object.
2) Transcript has start_not_found set if the translation doesn't begin with Methionine.
3) Transcript has end_not_found set if the translation doesn't end with a stop.
4) The correct selenocysteine remark and coordinates are automatically added if "seleno" appears in an annotation remark for the transcript.
5) Locus has a description (also known as "full name").
6) Transcripts within each locus are all on the same strand.
7) Transcripts do not have a 5' UTR with start_not_found of 1, 2 or 3. (UTR start_not_found has been added as a menu option.)
8) There is evidence attached to each transcript.
9) Nucleotide evidence is only used once in each locus.
10) The same locus name root is not used for transcript names in more than one locus.
11) All the transcript names in the same locus have the same locus name root.
12) Transcript names start with the locus name if the locus name ends with "dot-number" (which means the clonename in such circumstances).
13) Transcript names end "dash-digit-digit-digit".

Zmap User Manual

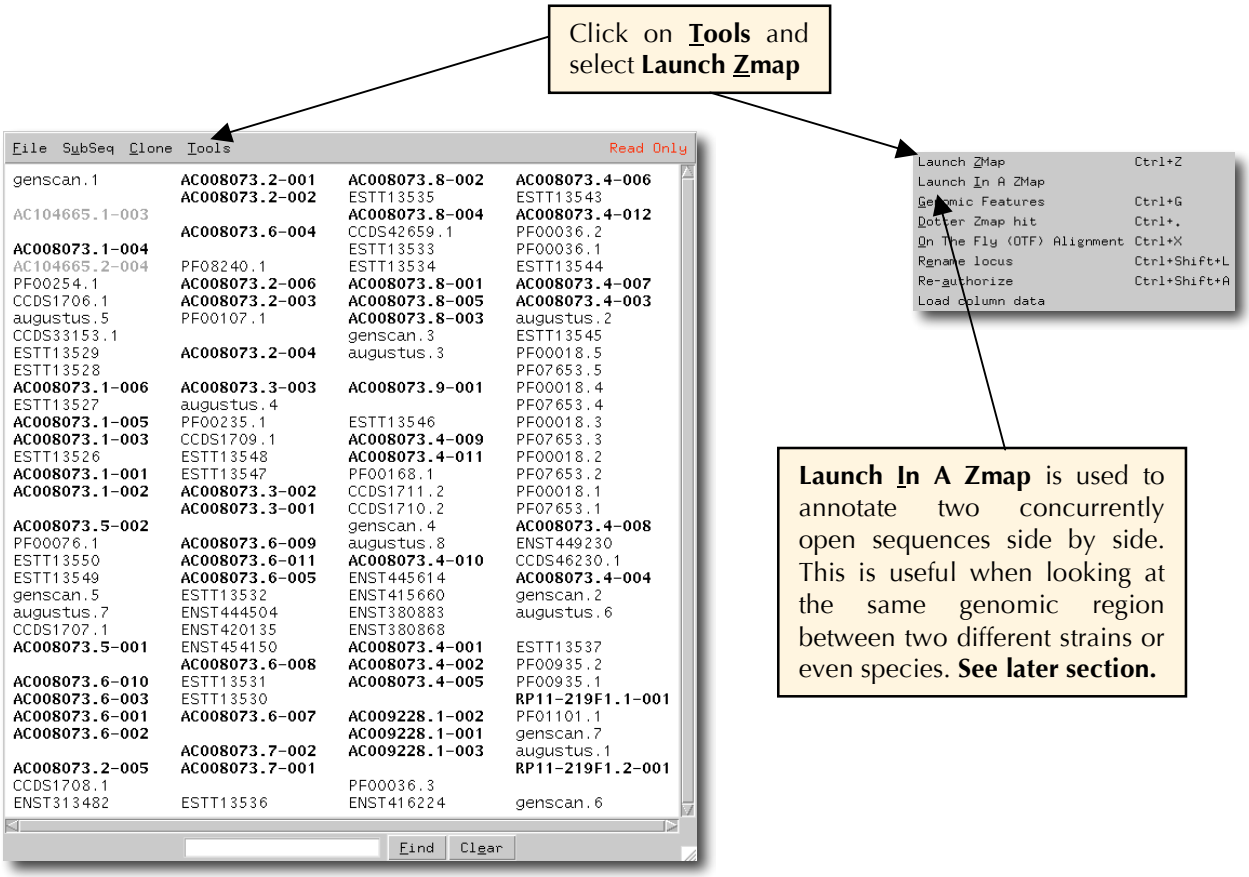
Written by Charles Steward and Laurens WIlming
(cas@sanger.ac.uk, lw2@sanger.ac.uk)
Wellcome Trust Sanger Institute

Zmap

Zmap is a software package that provides a visualisation tool for genomic features. The software is written in C, utilising the gnome toolkit (GTK2) to draw features on a canvas. Zmap accepts input from multiple sources in multiple formats across multiple genomes and is written in a way so that the addition of further formats is made as trivial as possible. Currently the list of formats includes GFF and DAS, which may reside in any one of; a file, an acedb instance, an http server. Multiple genomes and their associated features can be displayed in a single view as aligned blocks providing support for comparative annotation. Zmap does not include any utility for editing the features that it displays. It does however provide a powerful external interface with which to modify the features displayed on the canvas. Using this interface, Otterlace is used to annotate sequences present in the Otter database. This in turn updates to the Vertebrate Genome Annotation (VEGA) website (http://vega.sanger.ac.uk/index.html)

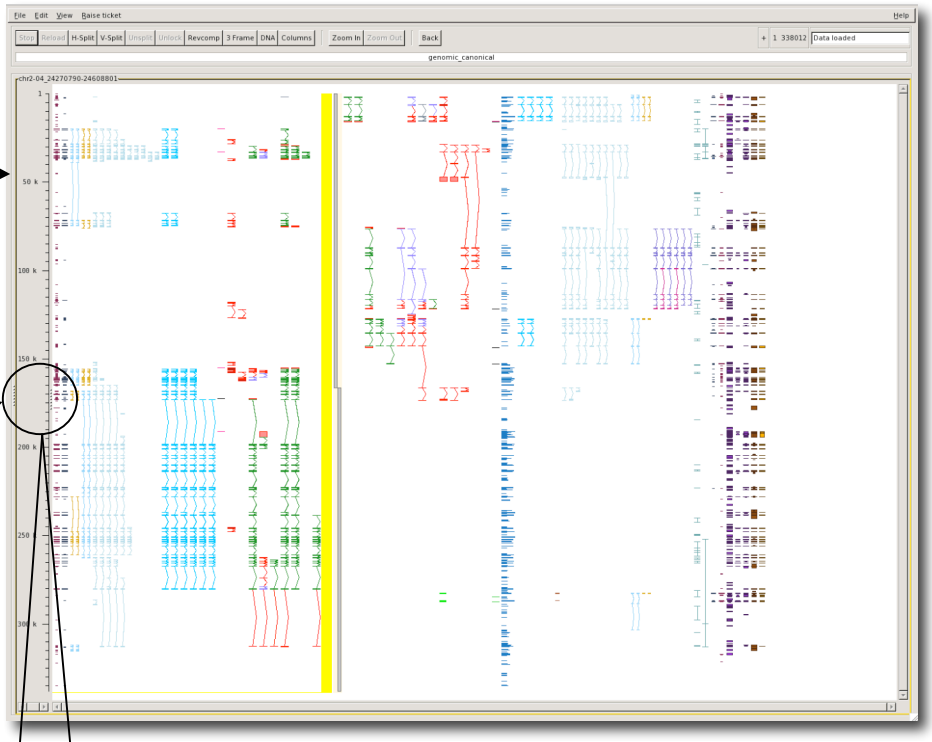
Opening Zmap

Zmap is opened via the Tools menu bar in Otterlace.


Click on **Tools** and select **Launch Zmap**
File SubSeq Clone Tools
Read Only
Launch ZMap Ctrl+Z
Launch In A ZMap
Genomic Features Ctrl+G
Dotter Zmap hit Ctrl+.
On The Fly (OTF) Alignment Ctrl+X
Rename locus Ctrl+Shift+L
Re-authorize Ctrl+Shift+A
Load column data
Find
Clear
Launch In A Zmap is used to annotate two concurrently open sequences side by side. This is useful when looking at the same genomic region between two different strains or even species. **See later section.**

Main Zmap interface

This is the main **Zmap** interface showing an overview of any analysis and annotation that may be present in your region of interest.

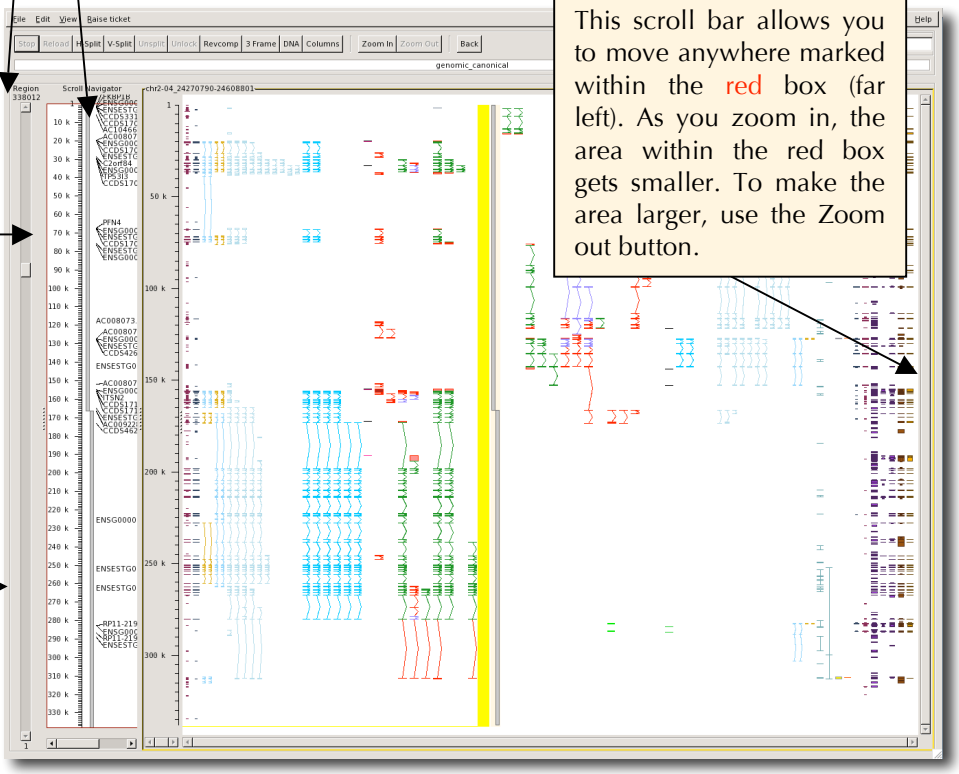


There are various hidden options that you can reveal by dragging the **dotted** regions.

The red box shows the extent of the sequence displayed in the main window showing the analysis, any previously annotated loci or any imported genes that are present in the clone.

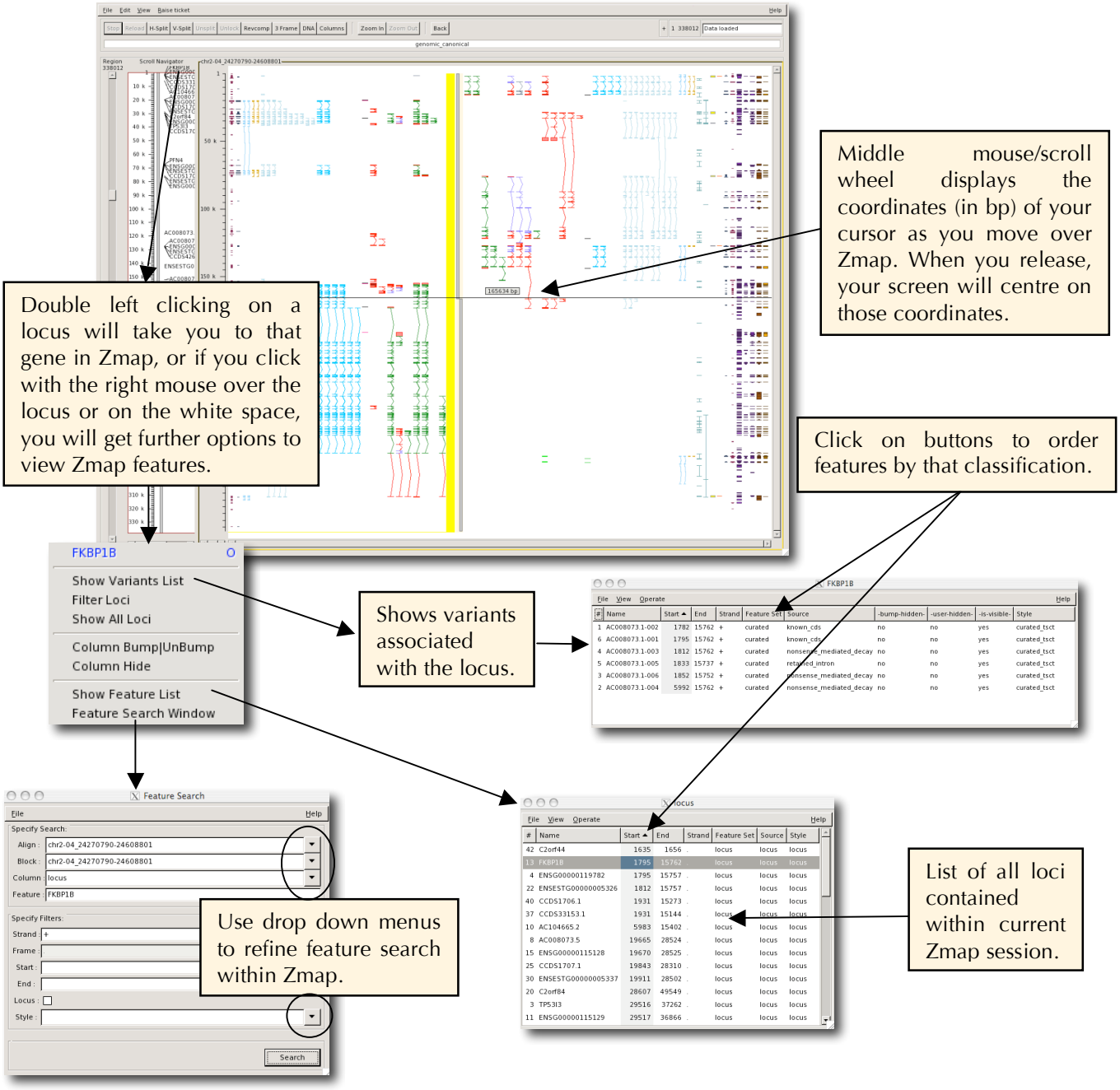
This panel has a scroll bar to show you where you are within the chromosome. It will allow you to jump to different regions. It is generally only useful if you open up very large sections of a chromosome.

This scroll bar allows you to move anywhere marked within the red box (far left). As you zoom in, the area within the red box gets smaller. To make the area larger, use the Zoom out button.



Navigating in Zmap and zooming options

1) Navigate by using the scroll bars or the middle mouse button. By clicking the middle mouse anywhere in Zmap you will see a horizontal line. You can move this up and down and the relative position in bp will be displayed along the line. When the button is released, the window will refresh, centering on the position of the line. You can also click in the window to make it active and use the scroll wheel to navigate up and down or achieve the same result using the scroll bar on the right hand side of the window. If you release the mouse outside the Zmap window, you can then check the sequence position displayed, without re-centering.

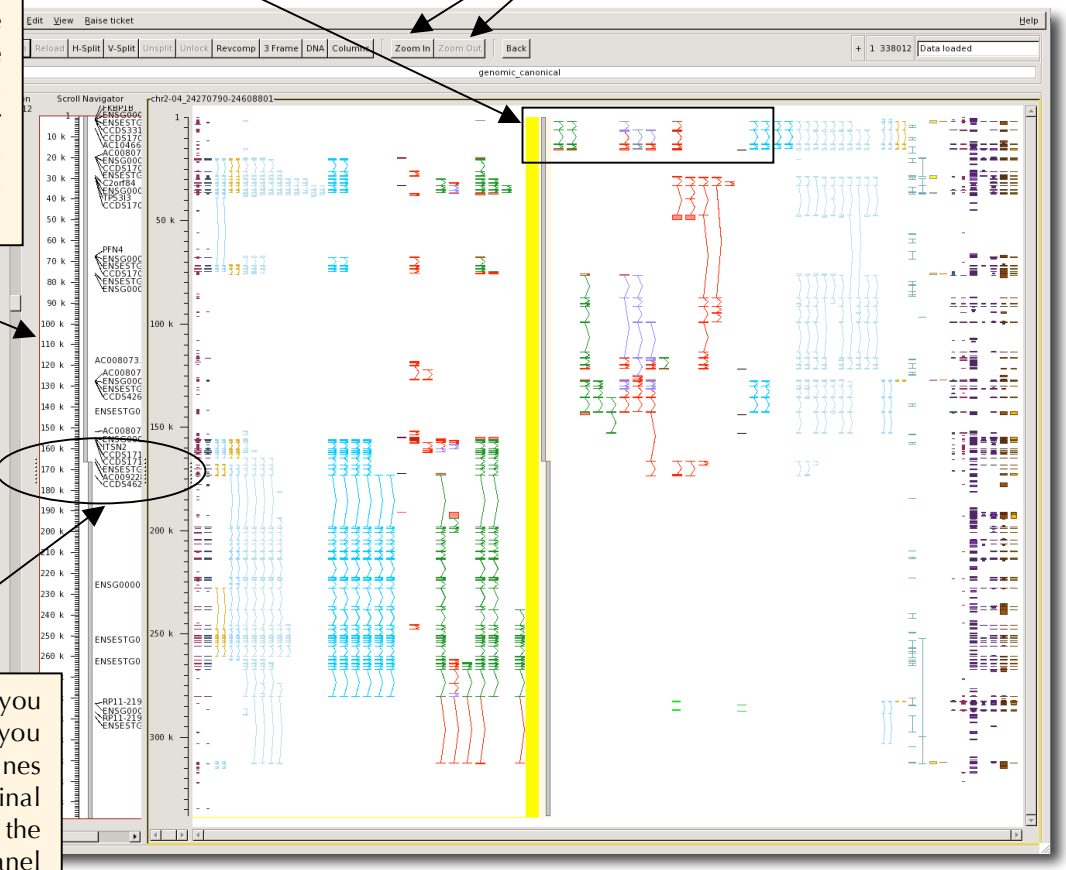

Middle mouse/scroll wheel displays the coordinates (in bp) of your cursor as you move over Zmap. When you release, your screen will centre on those coordinates.
Double left clicking on a locus will take you to that gene in Zmap, or if you click with the right mouse over the locus or on the white space, you will get further options to view Zmap features.
Click on buttons to order features by that classification.
FKBP1B
Show Variants List
Filter Loci
Show All Loci
Column Bump|UnBump
Column Hide
Show Feature List
Feature Search Window
Shows variants associated with the locus.
Feature Search
Use drop down menus to refine feature search within Zmap.
List of all loci contained within current Zmap session.

2) Zoom in by using the Zoom in/Zoom out buttons at the top, or by drawing a rectangle around the area of interest with the left mouse button. Use the "**z**" key on the keyboard to zoom to whatever feature is highlighted. Use the "**Z**" key to zoom to a whole transcript if you have an exon (s) highlighted or all HSPs if you have one HSP highlighted (HSPs are the "blocks" that you see in the homology columns, such as ESTs and protein hits).

To mark the rectangle click and hold the left mouse button at the top left of the area you want to outline and then drag out the outline until it encloses the area you want to zoom to. When you release the button, Zmap zooms in to that rectangle.

Use these buttons to Zoom in to a region or to Zoom out.

The red box is draggable. You can use the left mouse to alter the bounds of the display in the main window and the scrollbar to the right of the main window to scroll through the data quickly.

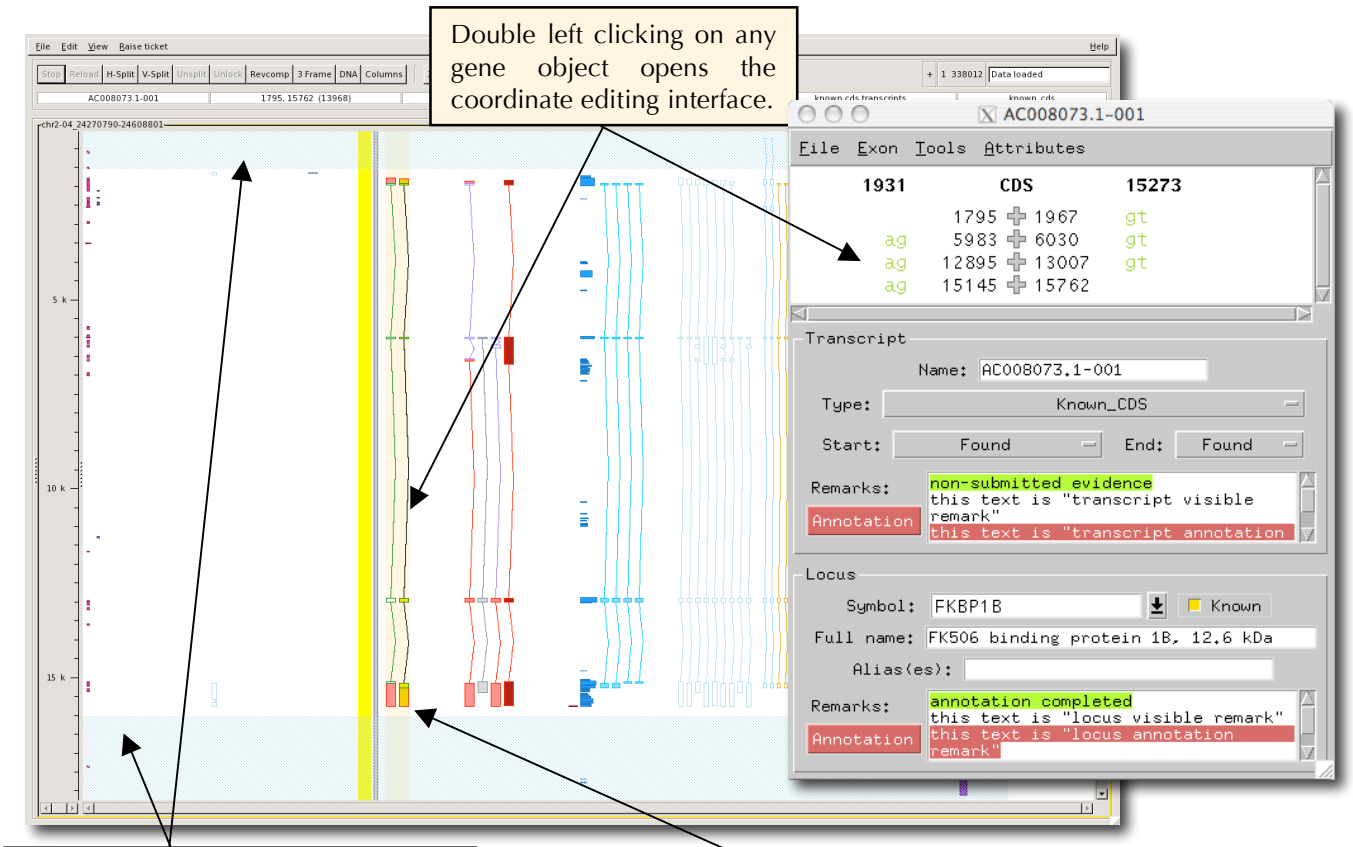


To save space when you are inspecting a region you can drag the dotted lines back to their original position to remove the scroll bar and locus panel information. Note, it is not necessary to have any of these panels open while you work.

The Focus Feature vs the Marked Feature

If you click on a column background then that column becomes the "focus" column and you can do various short cut operations on it such as pressing "b" to bump it. If you click on a feature then that feature becomes the "focus" feature and similarly you can do various short cut operations on it such as zooming in to it. (Note when you select a feature then its column automatically becomes the focus column.)

While the focus facility is useful, the focus changes every time you click on a new feature. Sometimes you want to select a "working" feature or area more permanently. To do this you can "mark" the feature or area and it will stay "marked" until you unmark it. "Marking" an area within Zmap to work on is essential, allowing you to work much faster. The "marked" area is left clear while the unmarked area above and below is marked with a blue overlay (see screen shot below):


Double left clicking on any gene object opens the coordinate editing interface.
AC008073.1-001
File Exon Tools Attributes
1931 CDS 15273
1795 1967 gt
ag 5983 6030 gt
ag 12895 13007 gt
ag 15145 15762
Transcript
Name: AC008073.1-001
Type: Known_CDS
Start: Found End: Found
Remarks:
non-submitted evidence
this text is "transcript visible remark"
Annotation
this text is "transcript annotation
Locus
Symbol: FKBP1B
Known
Full name: FK506 binding protein 1B, 12.6 kDa
Alias(es):
Remarks:
annotation completed
this text is "locus visible remark"
Annotation
this text is "locus annotation remark"
The marked area is designated by the blue shading at the top and bottom of the screen shot. The boundaries can be manually changed – see next page on manual cropping.
This screen shot shows a column that has been selected and then marked.

Mark a feature

1) Select a feature to make it the focus feature.
2) Press "m" to mark the feature, the feature will be highlighted with a blue overlay.

Feature marking behaves differently according to the type of feature you highlighted prior to marking and according to whether you press "m" or "M" to do the marking:

1) If you press “m”, the mark is made around all features you have highlighted, e.g. a whole transcript, a single exon, several HSPs.
2) If you press "M" to do the marking around transcripts the whole transcript becomes the marked feature and the marked area extends from the start to the end of the transcript.
3) If you press "M" to do the marking around alignments all the HSPs for that alignment become the marked feature and the marked area extends from the start to the end of all the HSPs.
4) If you press "M" to do the marking around all other features: the feature becomes the marked feature and the marked area extends from the start to the end of the feature.
5) If no feature is selected but an area was selected using the left button rubberband then that area is marked.
6) If no feature or area is selected then the visible screen area minus a small top/bottom margin is marked.

Mark an area

1) Select an area by holding down the left mouse button and dragging out a box to focus on that area.
2) Press "m" to mark the area.

Manual cropping of the marked borders

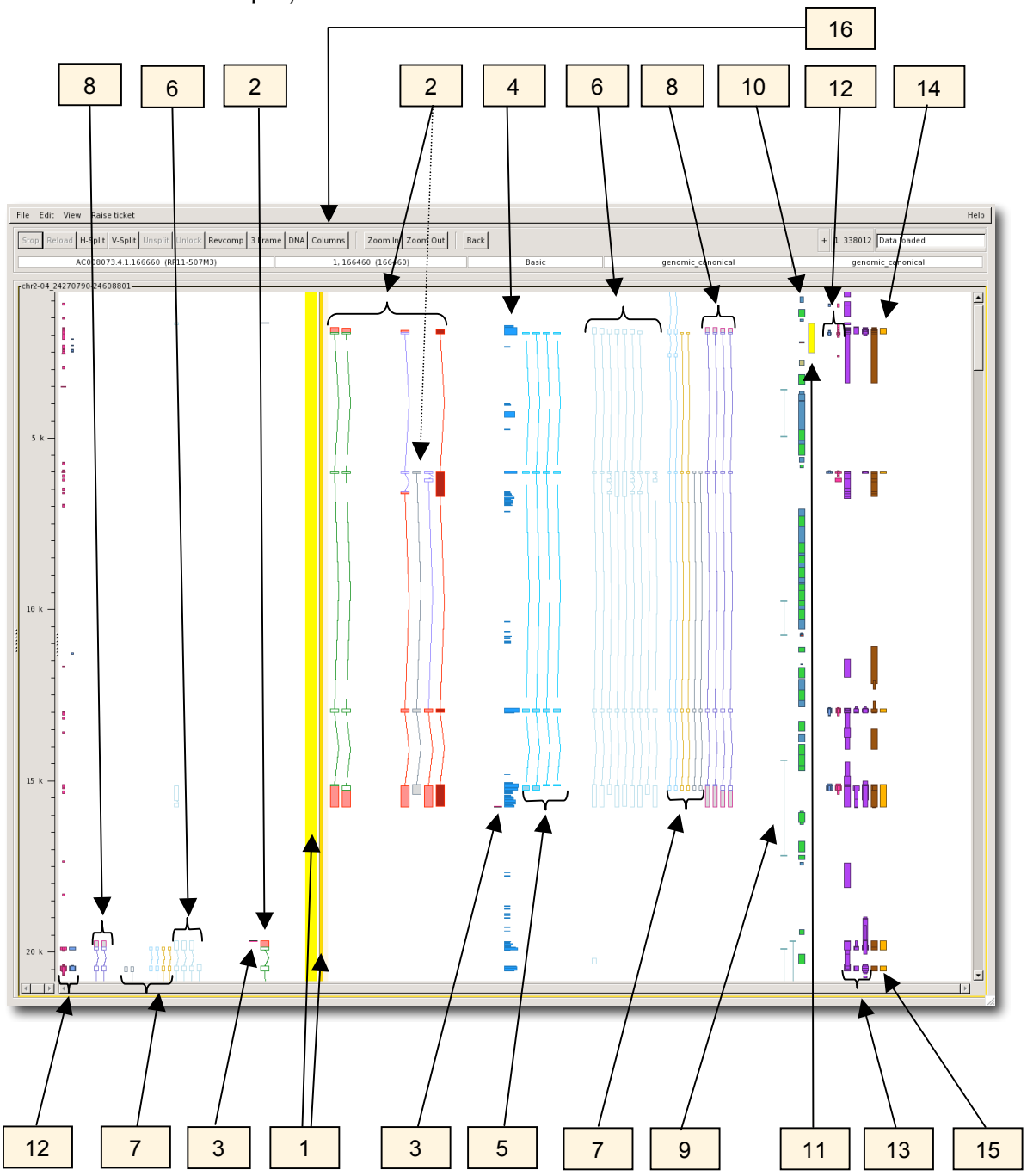
You can manually change the borders of the marked area by putting your cursor over this area and using the cropping tool by clicking and holding with the left mouse button and dragging to make the area bigger or smaller.

Unmark a feature

Press "m" or "M" again, i.e. the mark key toggles marking on and off.

General Zmap display features

Different features are displayed in distinct columns as follows:



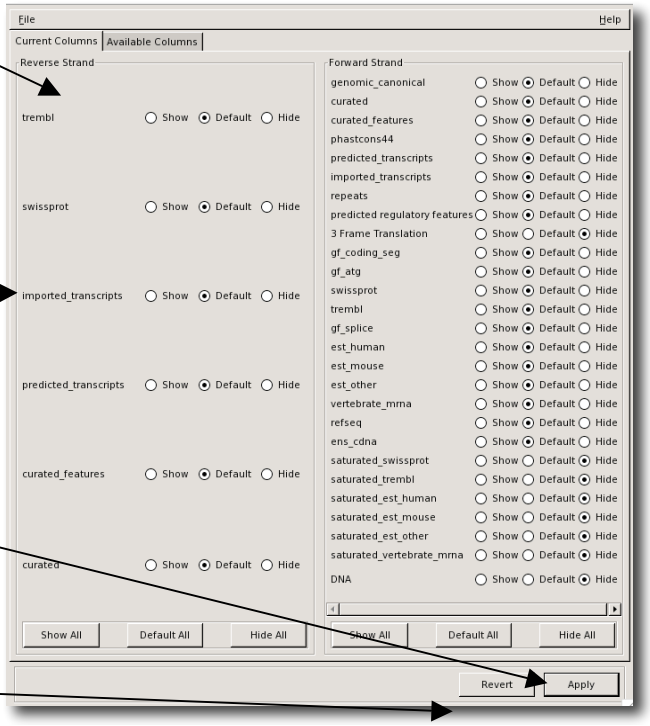
Note - you may see more or fewer features and columns depending on how your preferences are set up. For descriptions of all column types such as DAS sources, visit this URL -
http://scratchy.internal.sanger.ac.uk/wiki/index.php/Otterlace_filter_descriptions

1) The thick yellow line represents the genomic sequence; everything to the left represents the negative strand and everything to the right the positive strand. DNA matches (i.e. ESTs, mRNAs and RefSeq) and repeats are all displayed to the right of the center although they may align to either strand. The thin bar to the right is the clone that the genomic sequence is made up from. Double click on this to access the DE editing window.
2) Annotated transcripts; green is coding (CDS), red is non-coding (UTR and transcript variants) and purple shows the "coding" region of NMD variants. Grey transcripts (see dotted line) contain exons outside the sequence slice being viewed and should not be confused with Halfwise hits.
3) Curated features, such as PolyA features are seen as horizontal black lines.
4) Phastcons44 – conserved regions detected using multiple sequence alignments of 44 organisms.
5) Imported annotation from CCDS (human and mouse only).
6) Imported transcripts via DAS source. Here PASA_ESTs are shown.
7) Predicted transcripts such as Genscan (pale blue), Augustus (gold) and Halfwise predictions of Pfam (grey).
8) Imported annotation from Ensembl.
9) gis_pet_ditags and chip_pet_ditags are indicators of transcript boundaries.
10) Repeats (blue=Line , light green=Sine , gold=other), tandem repeats are red.
11) CpG islands appear as yellow boxes.
12) Protein matches are strand specific - SwissProt are light blue and Trembl pink.
13) EST matches are displayed as purple blocks and are broken down into human ESTs, mouse ESTs, and other ESTs from other organisms. 5' reads are on the left and 3' on the right.
14) mRNA matches contains all species and are displayed as brown blocks,
15) RefSeq matches are the orange blocks.
16) Features and analysis available

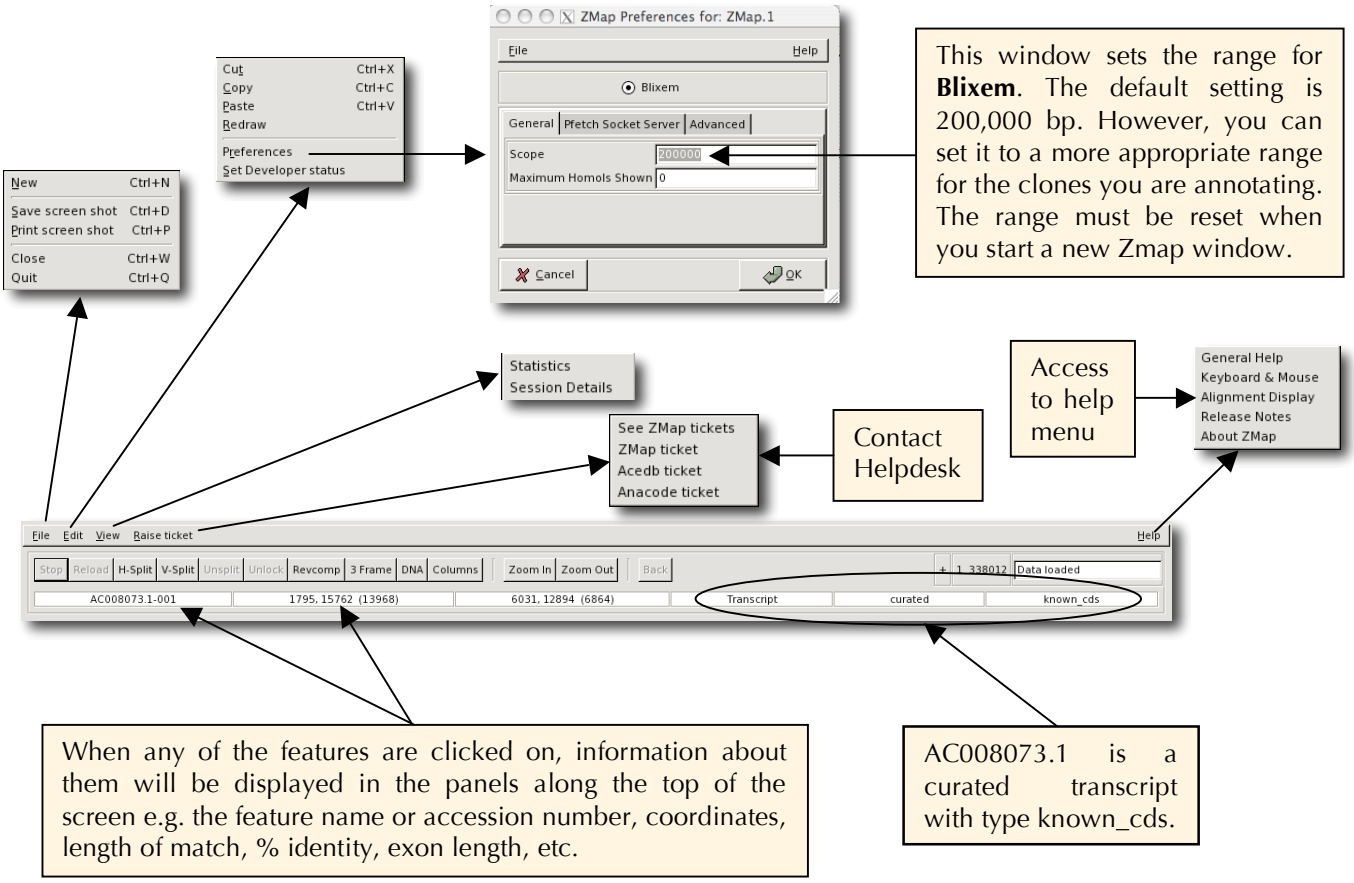
14) The **Columns** button brings up this window, allowing you to customize Zmap by turning features on and off.

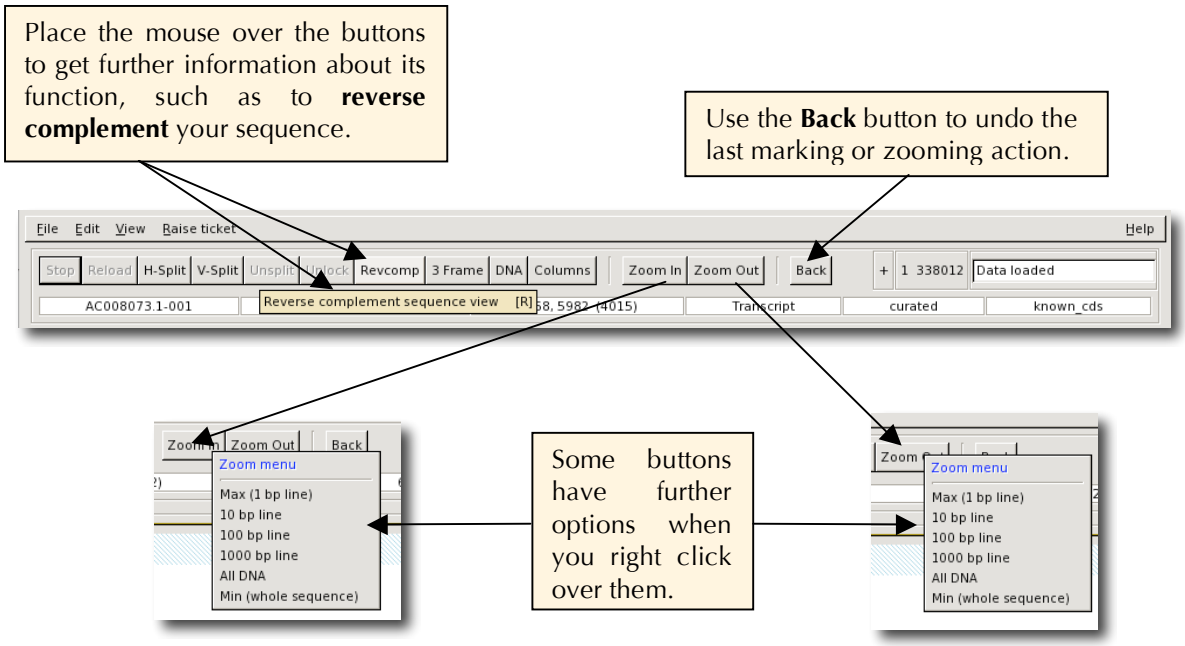
Select the features that you want to be visible on Zmap and click on **Apply**.

Revert sets the features to the default setting.

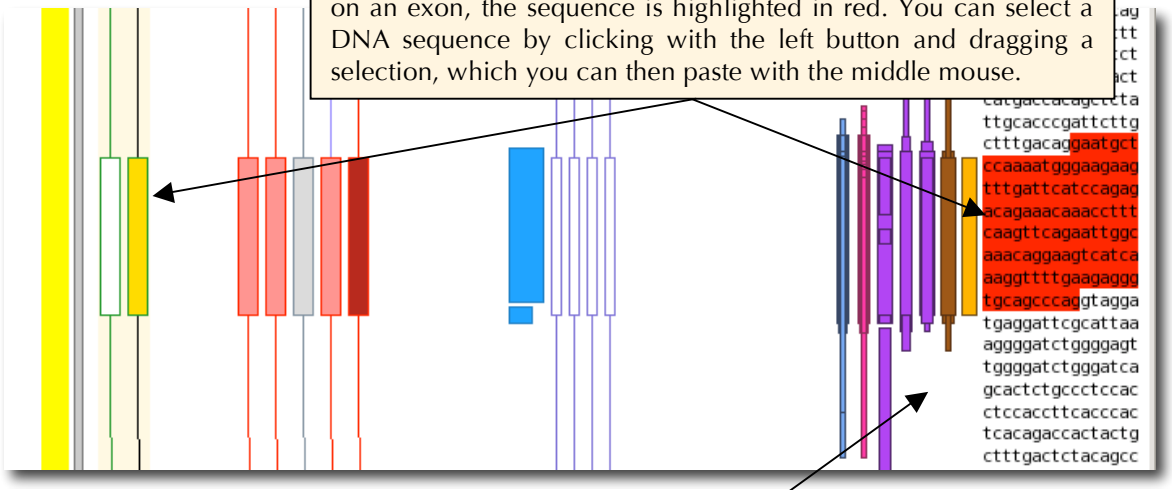


Functionality of the features at the top of the Zmap display.

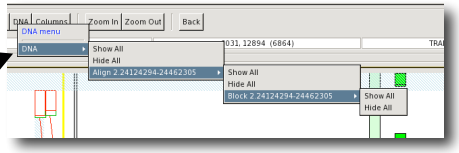


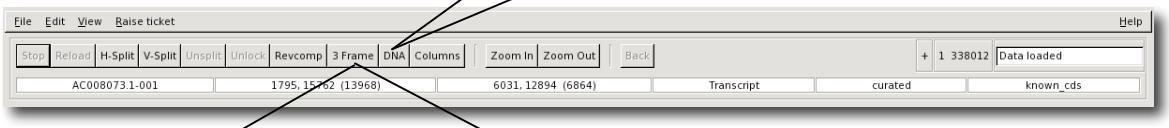

Place the mouse over the buttons to get further information about its function, such as to **reverse complement** your sequence.
Use the **Back** button to undo the last marking or zooming action.
File Edit View Raise ticket
Help
Stop Reload H-Split V-Split Unsplit Unlock Revcomp 3 Frame DNA Columns
Zoom In Zoom Out
Back
1 338012
Data loaded
AC008073.1-001
Reverse complement sequence view [R]
Transcript
curated
known_cds
Zoom menu
Max (1 bp line)
10 bp line
100 bp line
1000 bp line
All DNA
Min (whole sequence)
Some buttons have further options when you right click over them.

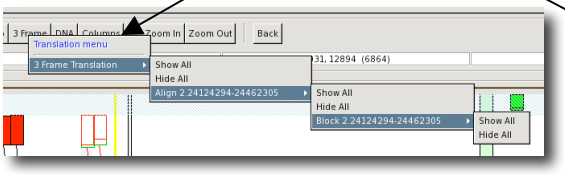
The **DNA** button will show the nucleotide sequence. If you click on an exon, the sequence is highlighted in red. You can select a DNA sequence by clicking with the left button and dragging a selection, which you can then paste with the middle mouse.

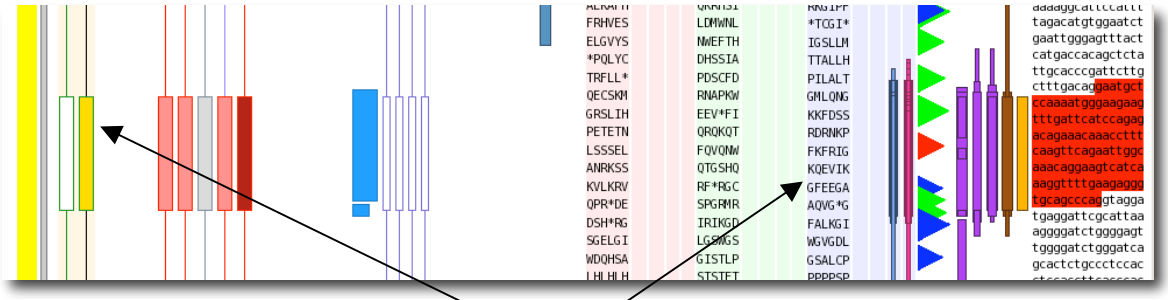


Click the buttons with the left mouse to operate the **DNA** and **3 Frame** translation options. Right click over the buttons for further options. To remove these displays from Zmap, click on the button again.




File Edit View Raise ticket Help
Stop Reload H-Split V-Split Unsplit Unlock Revcomp 3 Frame DNA Columns Zoom In Zoom Out Back
Data loaded
AC008073.1-001
Transcript
curated
known_cds

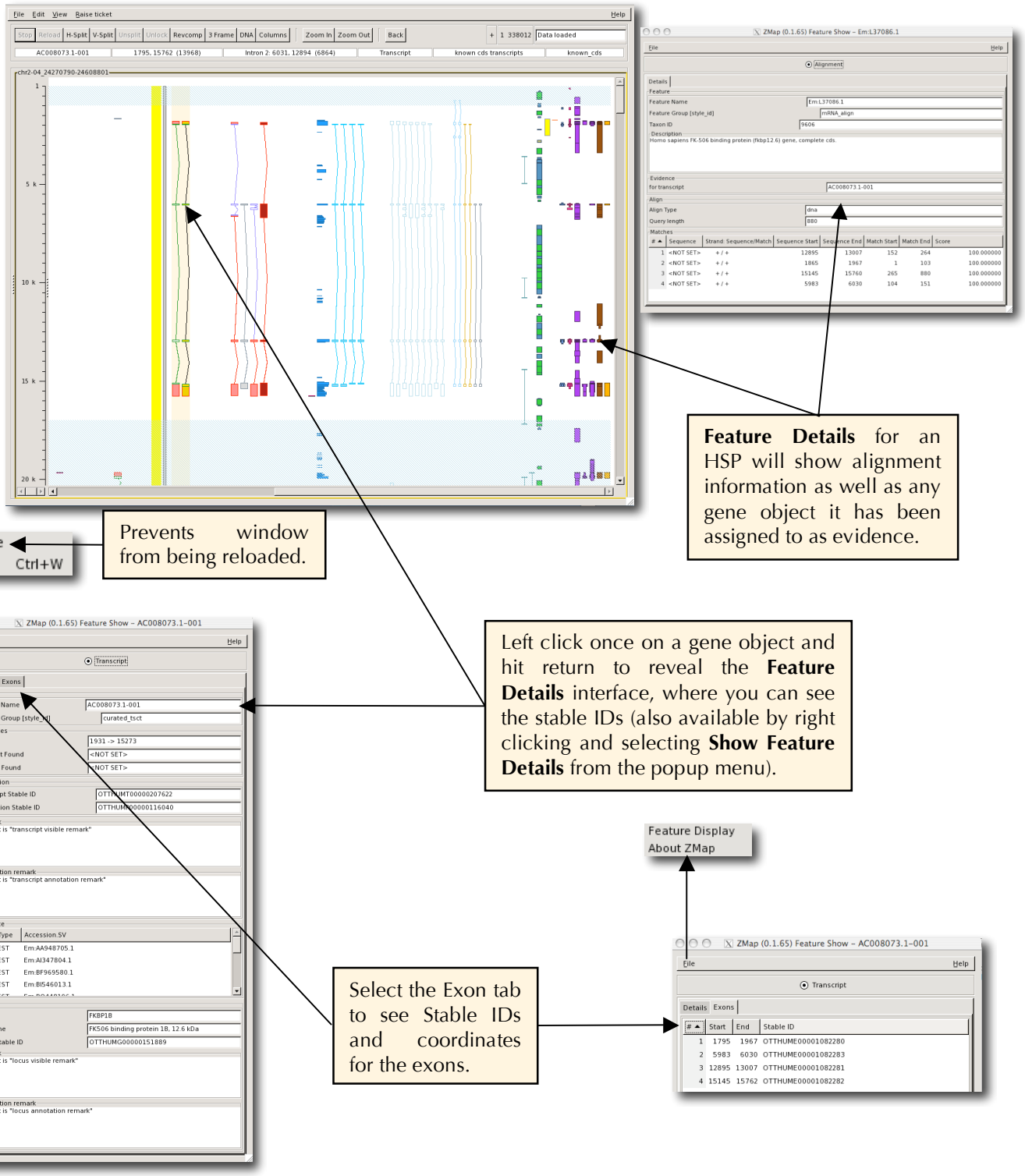




The **3 Frame** button will show the amino acid sequence in each of the three reading frames. If you click on an exon, the sequence is highlighted in red.

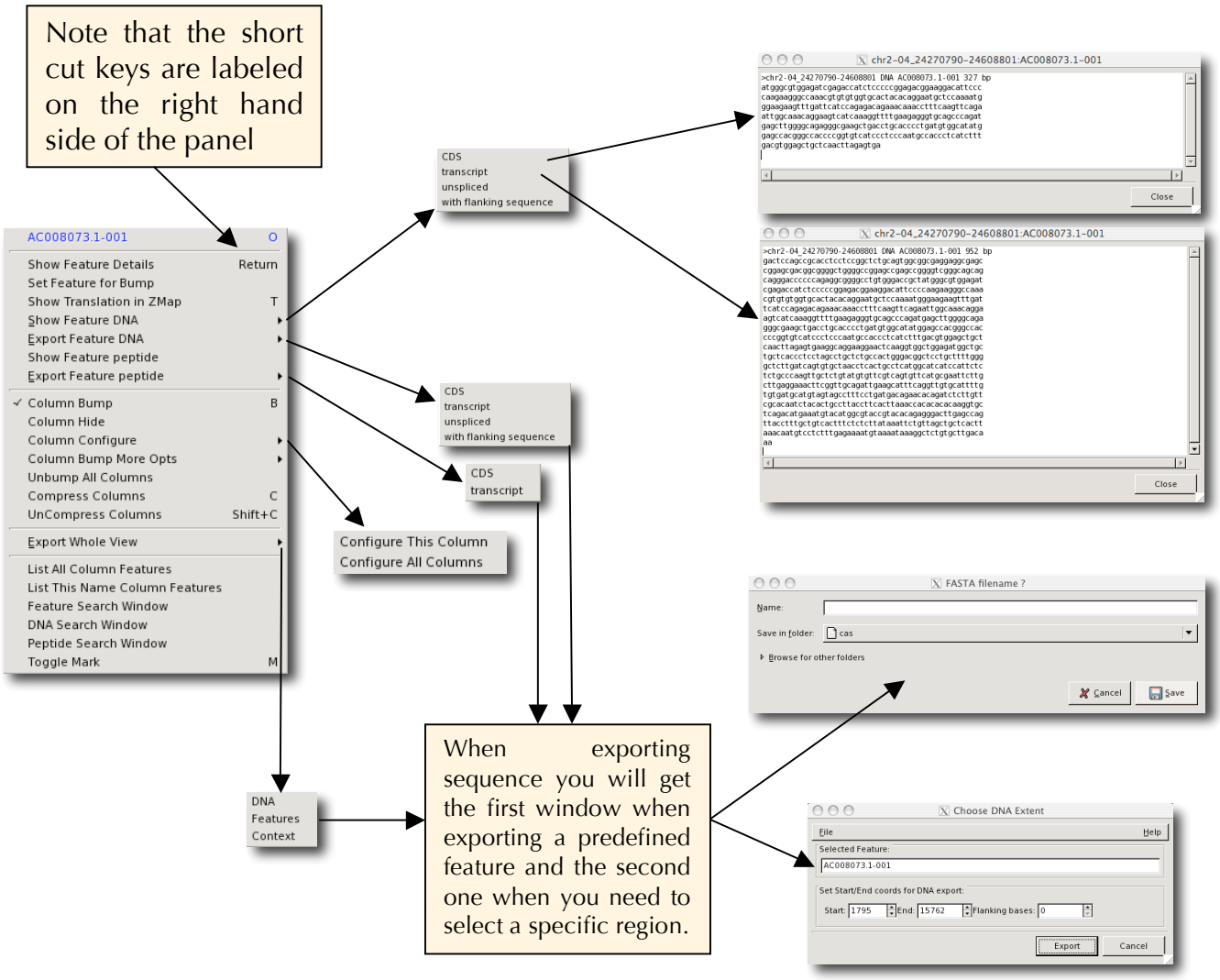
Show feature details

Right click on a gene object or ‘o’ key when highlighted to see information on otter IDs and Ensembl IDs. For BLAST hits, double click on the HSP to get the feature interface where you will find details on alignment and on what HAVANA object the HSP has been assigned to, if any:


Feature Details for an HSP will show alignment information as well as any gene object it has been assigned to as evidence.
Preserve
Close Ctrl+W
Prevents window from being reloaded.
Left click once on a gene object and hit return to reveal the Feature Details interface, where you can see the stable IDs (also available by right clicking and selecting Show Feature Details from the popup menu).
Feature Display
About ZMap
Select the Exon tab to see Stable IDs and coordinates for the exons.

Exporting features for gene objects

As described on the previous page, if you right click over any feature (or type "o" when a feature is highlighted) you get further information. These screen shots show how you can view and export an annotated sequence to your home directory in various different ways, such as dumping features directly. In the main Zmap window, right click on an annotated gene object. From the drop down menu select **Export Feature DNA** and choose sequence required from CDS, transcript, unspliced and with flanking sequence. Alternatively select **Export Feature peptide** and choose either CDS or transcript. Here you can see how to **Show Feature DNA** for annotated gene object AC008073.1-001 in FASTA format; firstly, the section of the transcript that corresponds to the CDS and secondly the whole transcript, including the untranslated region (UTR).



Bumping features

This section describes how to select a feature, mark it and then zoom in to it and examine evidence that overlaps that feature. The default setting for Zmap is to show HSPs drawn on top of each other. This saves space on the canvas making it easier to see the general features of the region of interest. The bump option allows you to see the HSPs as multiple alignments.

1. Click on the feature you are interested in (perhaps a transcript)
2. Mark it by pressing "m"
3. Zoom in to the feature by pressing either "**z**" or "**Z**" (as described previously).

Now when you bump an evidence column to look at matches that overlap the feature you will find that bumping is much faster because only those matches that overlap the feature get bumped and you also have fewer matches to look at. The quickest way to bump a column is:

1. Click on the column to select it.
2. Bump it by pressing "b" (if you press "b" again the column will be unbumped). If you have marked a feature then bumping is restricted to matches that overlap that feature, otherwise bumping is for the whole column.

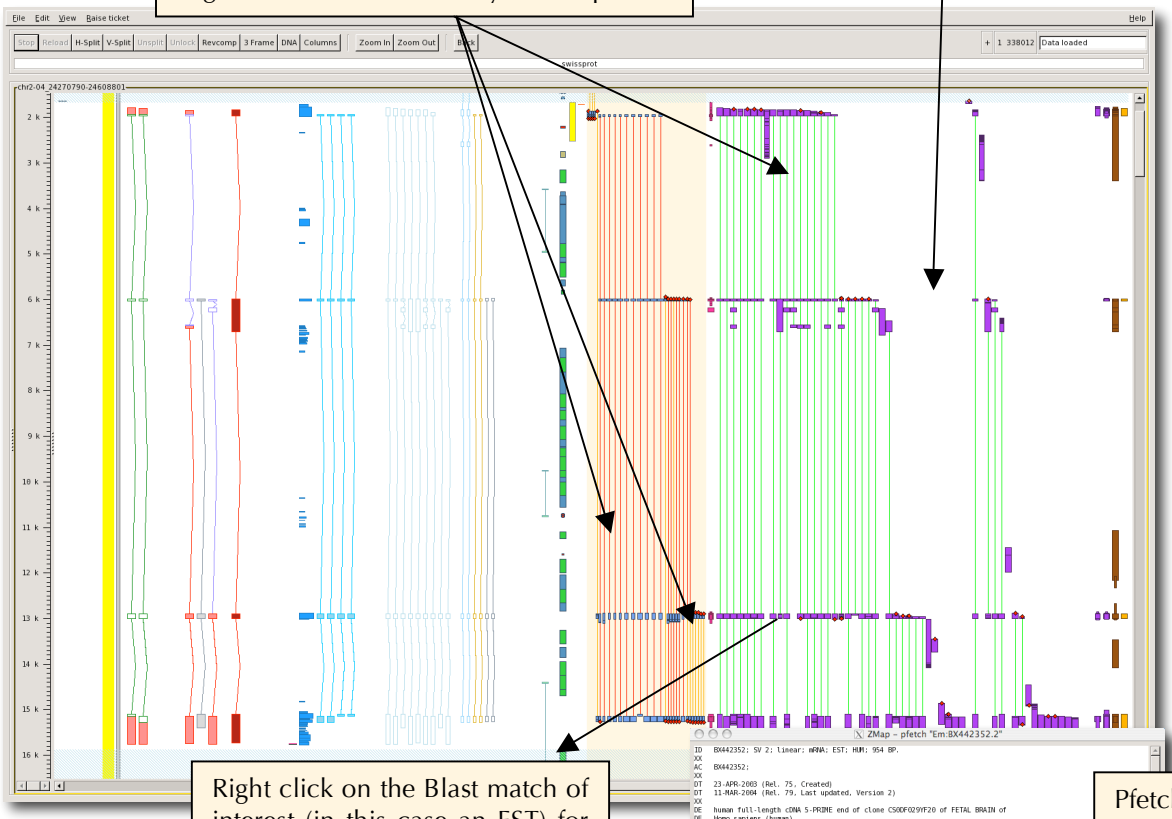
If you use the default bumping mode (i.e. you pressed "b") then you will find all matches from the same piece of evidence are joined by coloured bars, the colours indicate the level of colinearity between the matches (see next screen shot).

1. **Green**: the matches at either end are perfectly contiguous, e.g. 100, 230 ---> 231, 351
2. **Orange**: the matches at either end are colinear but not perfect, e.g. 100, 230 ---> 297, 351. Matches may also be this color when there are extra bases in the alignment, e.g. around clone boundaries.
3. **Red**: the matches are not colinear, e.g. 100, 230 ---> 141, 423

Alignment quality of the HSPs is depicted by the width of every alignment displayed since the width is a measure of that HSP's score. Therefore, the wider it is the closer the score is to 100%. The precise score is displayed in the Zmap details bar by clicking on the alignment. If HSPs are missing either the first or last Blast alignments in the set, they are marked with a red diamond at their start/end respectively. This indicates if they do not start at the first base/amino acid and/or do not end with the last base/amino acid of the alignment sequence. The screen shot below shows what options you get when you right click over a homology – note that you can also select an HSP and type "o". You also get further options such as retrieving the EMBL file for that homology using pfetch and starting **Blixem**, see later section (note, HSPs do not need to be bumped to use **Blixem**).

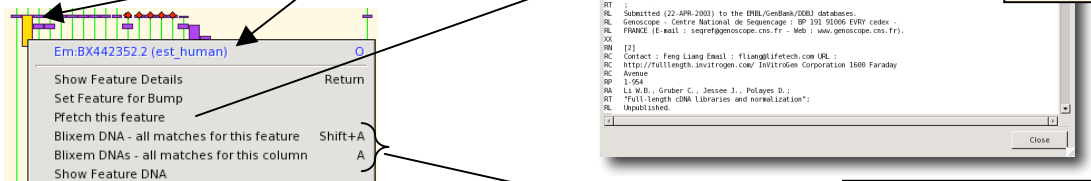
Note the different coloured lines for bumped homologies. The colouring allows you to see all matches for a piece of evidence instantly but also how good the alignment is for the feature you bumped.

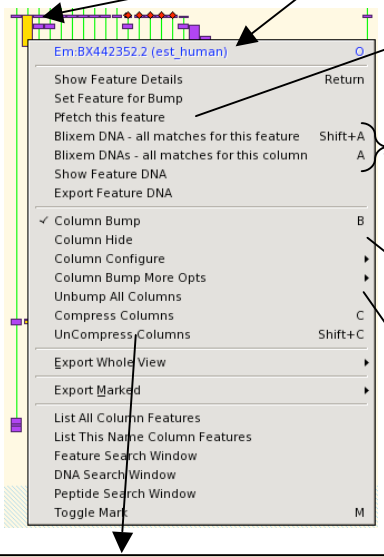
Note the red diamonds warning of missing sequence that cannot be aligned.



Right click on the Blast match of interest (in this case an EST) for more menu features.

Pfetch returns the EMBL flatfile for that sequence.





Allows you to inspect the sequence of just the chosen feature or all of the columns, aligned horizontally down to either the nucleotide or amino acid level against the genome. See later section on **Blixem**.

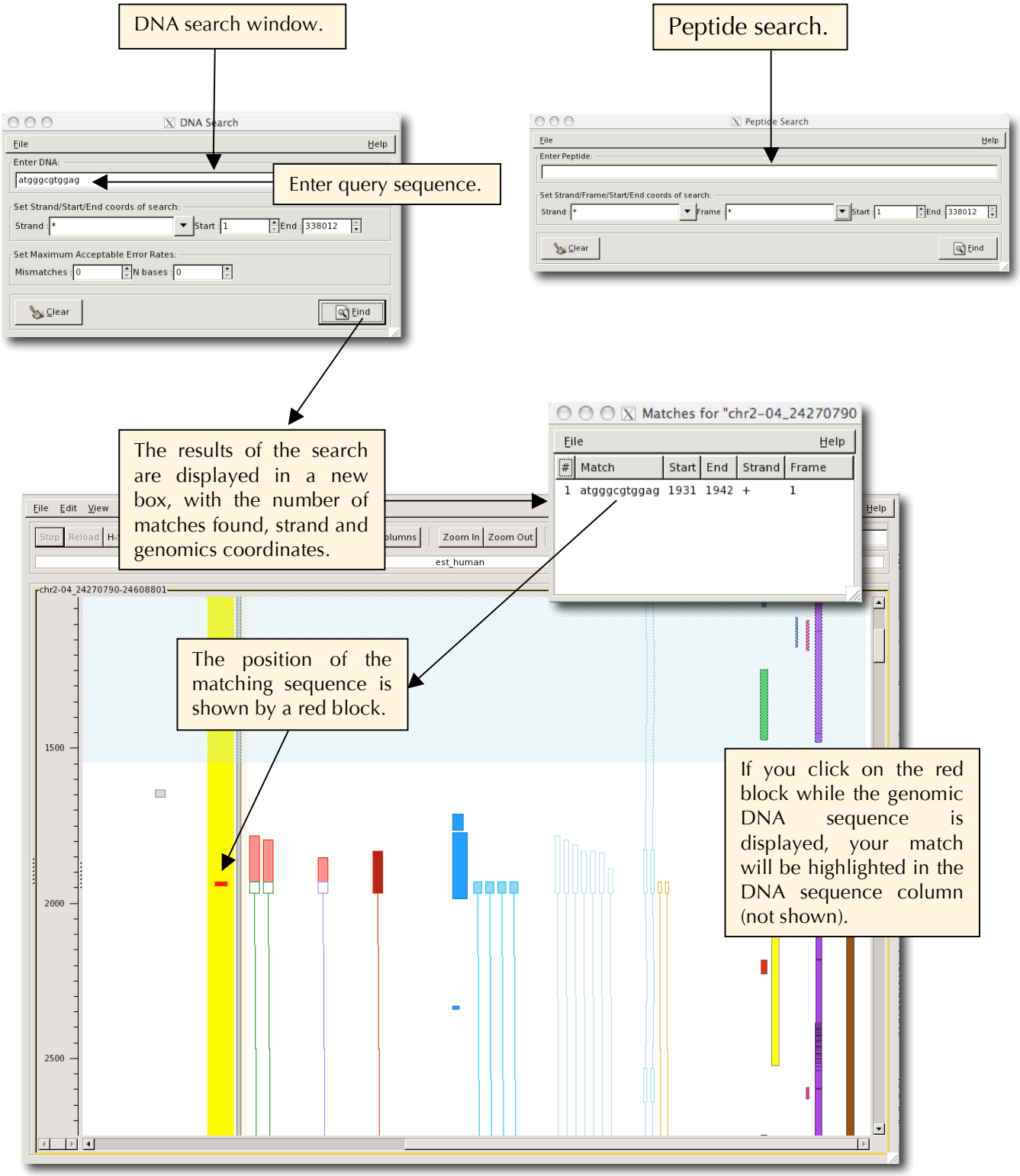
The ✓ shows that the column is bumped. Select it again to unbump it.

The Compress function removes excess white space by hiding columns that have no features in them, apart from those that have been set to “Show” in the “Columns” menu.

This menu allows you to change the way that bumping is displayed. There are multiple bump options, but the default is the most useful.

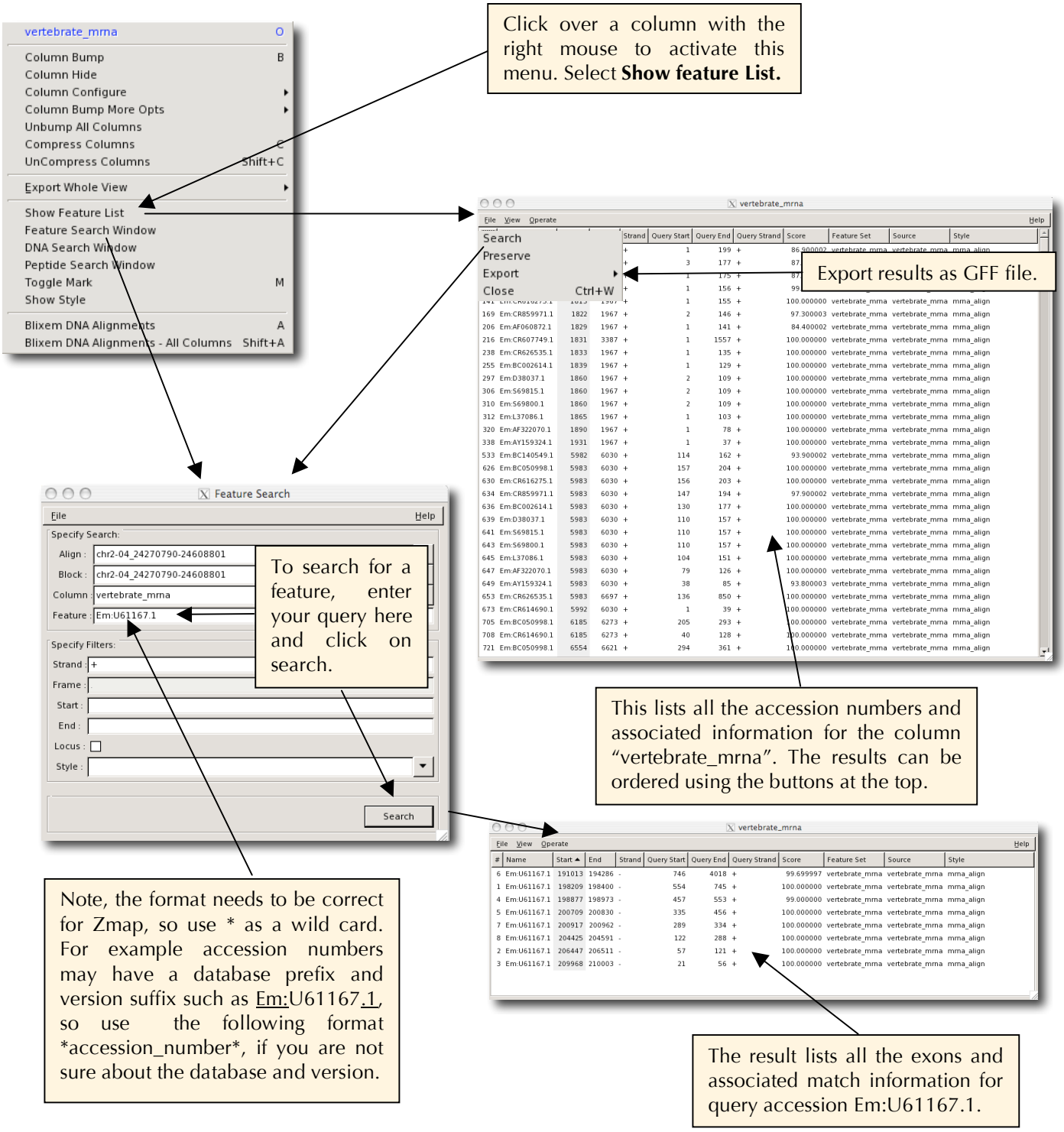
Searching for a sequence in Zmap

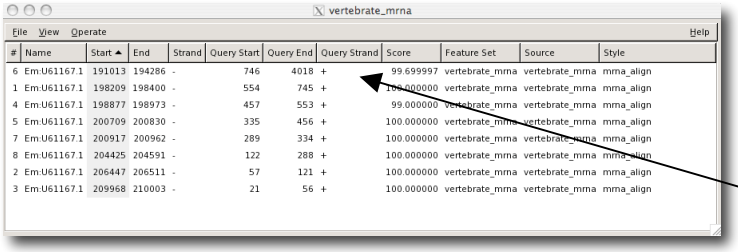
DNA and peptide search windows are provided from within Zmap and can be accessed by right clicking on Zmap space and selecting the option at the bottom of the menu. Both search windows are shown below:



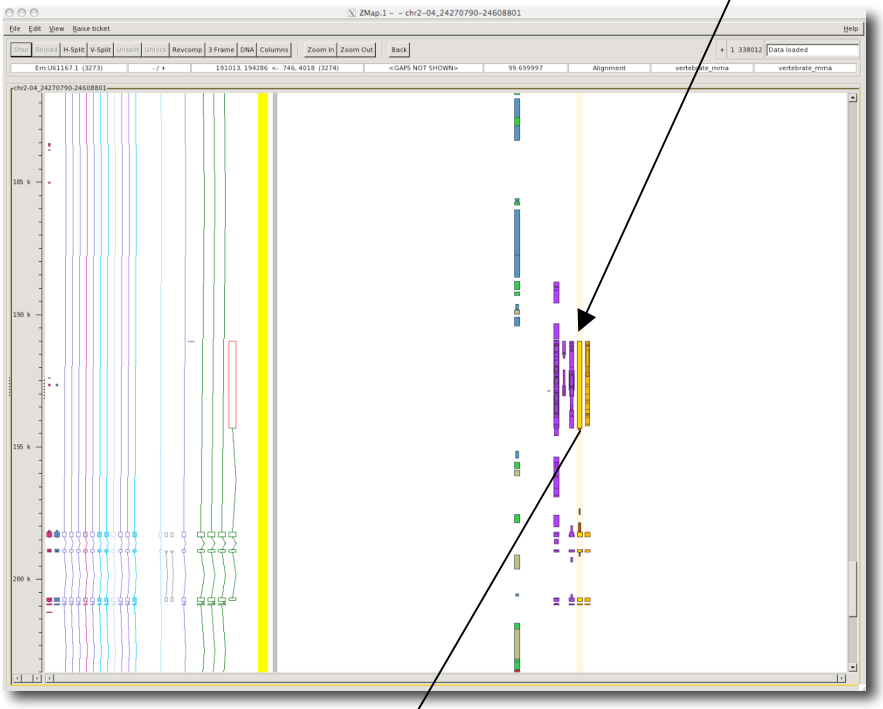
Searching for a feature in Zmap

This option allows you to list all the features contained in a column in one window. There are further options for you to search within these results to find a specific feature. The list of column features can be exported as a GFF file via the File menu.

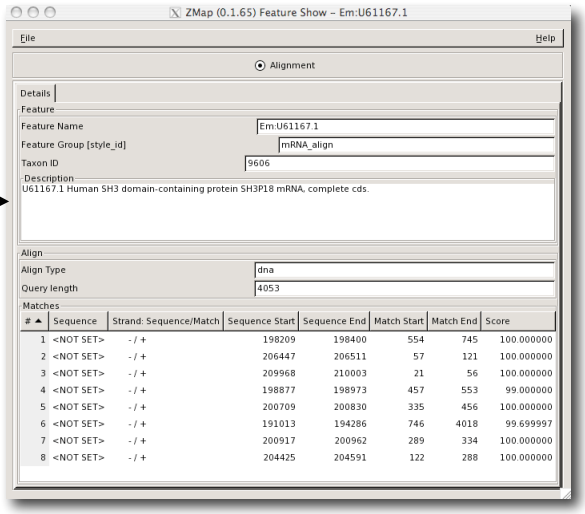

vertebrate_mrna
Column Bump
Column Hide
Column Configure
Column Bump More Opts
Unbump All Columns
Compress Columns
UnCompress Columns
Export Whole View
Show Feature List
Feature Search Window
DNA Search Window
Peptide Search Window
Toggle Mark
Show Style
Blixem DNA Alignments
Blixem DNA Alignments - All Columns
Click over a column with the right mouse to activate this menu. Select Show feature List.
Search
Preserve
Export
Close
Export results as GFF file.
Feature Search
Specify Search:
Align :
Block :
Column :
Feature :
Specify Filters:
Strand :
Frame :
Start :
End :
Locus :
Style :
Search
To search for a feature, enter your query here and click on search.
This lists all the accession numbers and associated information for the column "vertebrate_mrna". The results can be ordered using the buttons at the top.
Note, the format needs to be correct for Zmap, so use * as a wild card. For example accession numbers may have a database prefix and version suffix such as Em:U61167.1, so use the following format *accession_number*, if you are not sure about the database and version.
The result lists all the exons and associated match information for query accession Em:U61167.1.


vertebrate_mrna
File View Operate Help
Name Start End Strand Query Start Query End Query Strand Score Feature Set Source Style
6 Em:U61167.1 191013 194286 - 746 4018 + 99.699997 vertebrate_mrna vertebrate_mrna mrna_align
1 Em:U61167.1 198209 198400 - 554 745 + 100.000000 vertebrate_mrna vertebrate_mrna mrna_align
4 Em:U61167.1 198877 198973 - 457 553 + 99.000000 vertebrate_mrna vertebrate_mrna mrna_align
5 Em:U61167.1 200709 200830 - 335 456 + 100.000000 vertebrate_mrna vertebrate_mrna mrna_align
7 Em:U61167.1 200917 200962 - 289 334 + 100.000000 vertebrate_mrna vertebrate_mrna mrna_align
8 Em:U61167.1 204425 204591 - 122 288 + 100.000000 vertebrate_mrna vertebrate_mrna mrna_align
2 Em:U61167.1 206447 206511 - 57 121 + 100.000000 vertebrate_mrna vertebrate_mrna mrna_align
3 Em:U61167.1 209968 210003 - 21 56 + 100.000000 vertebrate_mrna vertebrate_mrna mrna_align

If you now left double click on the match you want to inspect, Zmap will zoom straight to it. Note, this may not work if you are searching for a feature out side of an area that is actively marked.



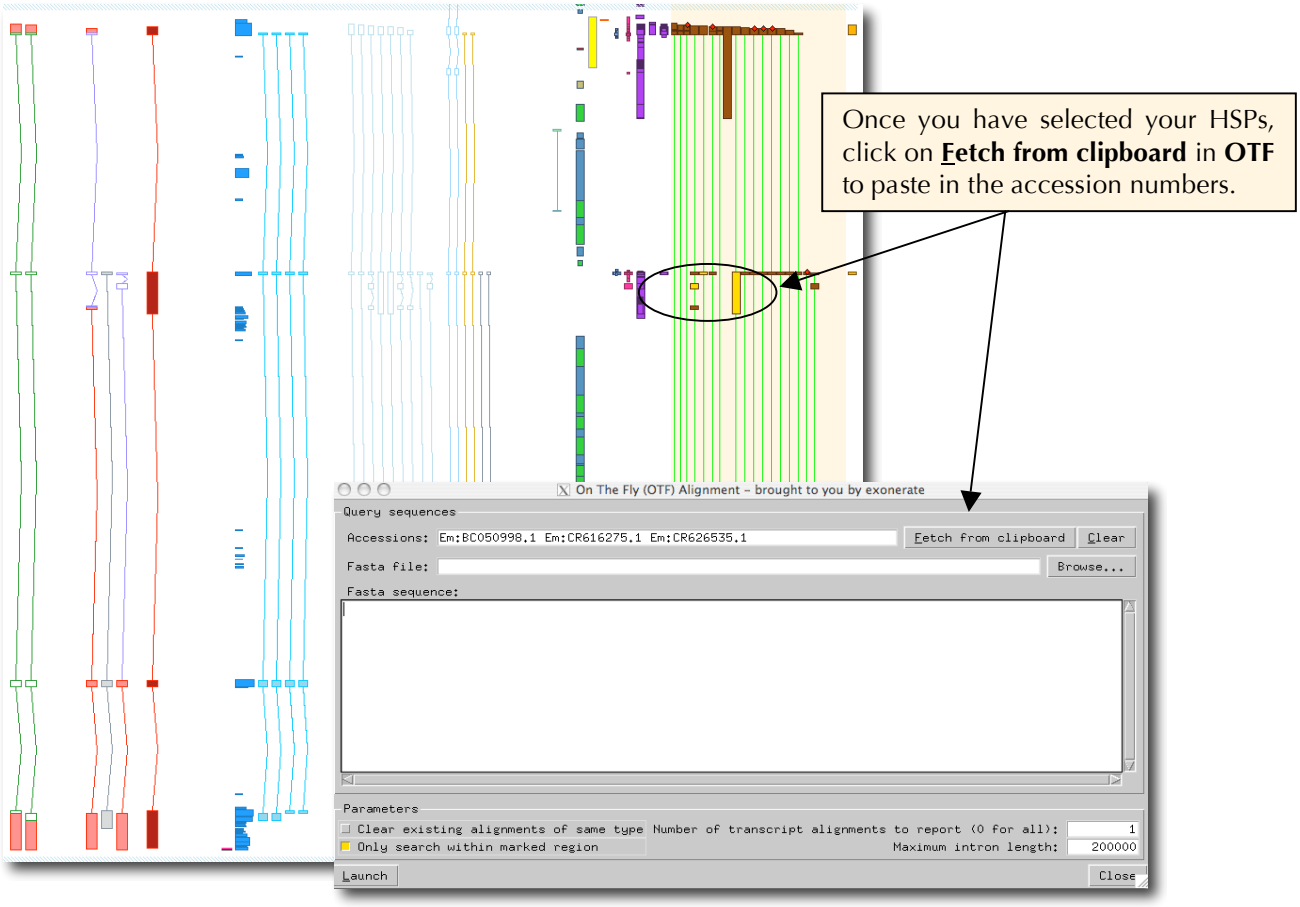
A further window will appear containing information about the feature.


ZMap (0.1.65) Feature Show - Em:U61167.1
File Help
Alignment
Details
Feature
Feature Name Em:U61167.1
Feature Group [style_id] mRNA_align
Taxon ID 9606
Description
U61167.1 Human SH3 domain-containing protein SH3P18 mRNA, complete cds.
Align
Align Type dna
Query length 4053
Matches
Sequence Strand: Sequence/Match Sequence Start Sequence End Match Start Match End Score
1 <NOT SET> - / + 198209 198400 554 745 100.000000
2 <NOT SET> - / + 206447 206511 57 121 100.000000
3 <NOT SET> - / + 209968 210003 21 56 100.000000
4 <NOT SET> - / + 198877 198973 457 553 99.000000
5 <NOT SET> - / + 200709 200830 335 456 100.000000
6 <NOT SET> - / + 191013 194286 746 4018 99.699997
7 <NOT SET> - / + 200917 200962 289 334 100.000000
8 <NOT SET> - / + 204425 204591 122 288 100.000000

Selecting single or multiple features and hiding/showing them

1) If you left click once on a feature in Zmap, you will highlight all of its exons, the coordinates of which are now stored in the paste buffer and can be copied elsewhere, such as into the transcript editing window in Otterlace.

2) You can select multiple features by holding the Shift key down and left clicking with mouse (same as for multi select on the Mac, Windows etc). This option will highlight a single exon at a time for each feature, but the accession numbers of each feature and the individual exon coordinates are held in the paste buffer. This is a particularly useful way of selecting Zmap hits to use in the OTF alignment tool, as all selected homologies will be held in the paste buffer and automatically pasted into the OTF accession window. Each of the exon coordinates can also be pasted into the transcript editing window in Otterlace.

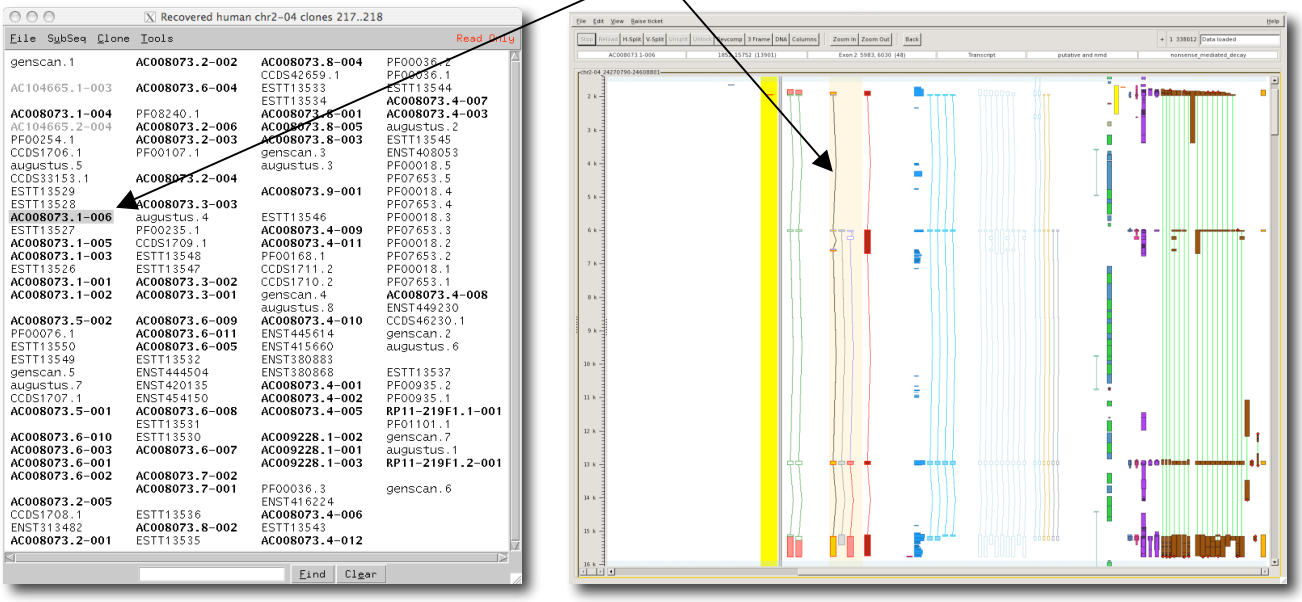

Once you have selected your HSPs, click on **Fetch from clipboard** in **OTF** to paste in the accession numbers.
On The Fly (OTF) Alignment - brought to you by exonerate
Query sequences
Accessions: Em:BC050998.1 Em:CR616275.1 Em:CR626535.1
Fetch from clipboard
Clear
Fasta file:
Browse...
Fasta sequence:
Parameters
Clear existing alignments of same type
Only search within marked region
Number of transcript alignments to report (0 for all): 1
Maximum intron length: 200000
Launch
Close

3) You can remove selected features in Zmap by pressing **Delete** on the keyboard and restore them by pressing **Shift-Delete** (note on the Mac you need to press **Fn-Delete** and **Shift-Fn-Delete**). This is a particularly useful way of removing evidence that you have already assigned to a transcript object.

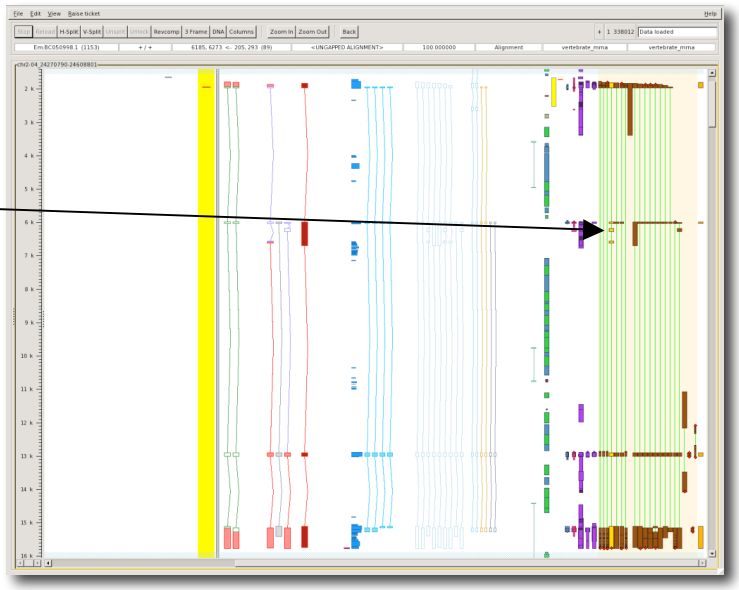
Rapid variant construction

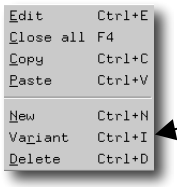
Otterlace and Zmap can be used together to generate variant objects quickly. Existing transcript objects can be used as a template for a new object while a Zmap HSP can be used to provide the coordinates for the new variant. The new object will take its transcript type from the parent.

1) Select the object that will form the foundation to the new variant, either by highlighting the object in Otterlace or clicking on the object in Zmap.

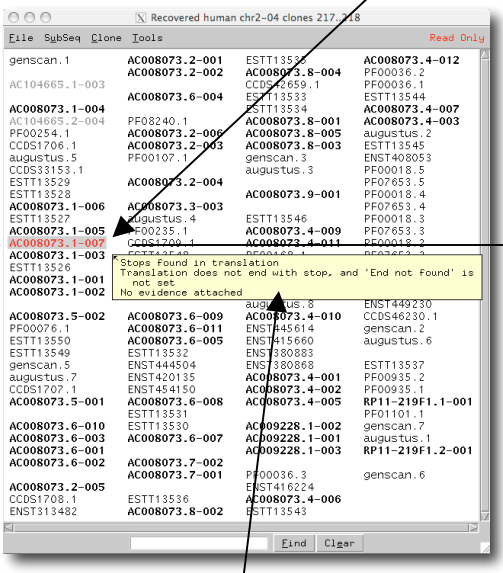


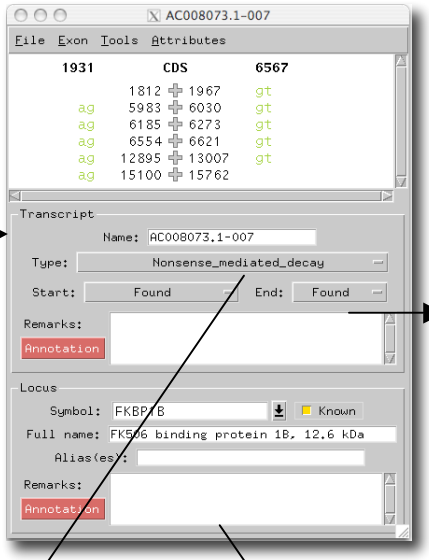
2) Click on the HSP that will give its coordinates to the new variant object.



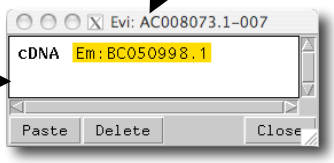


3) Now either use the key-stroke short cut or click on Variant. You will see a new object appear in your main window.

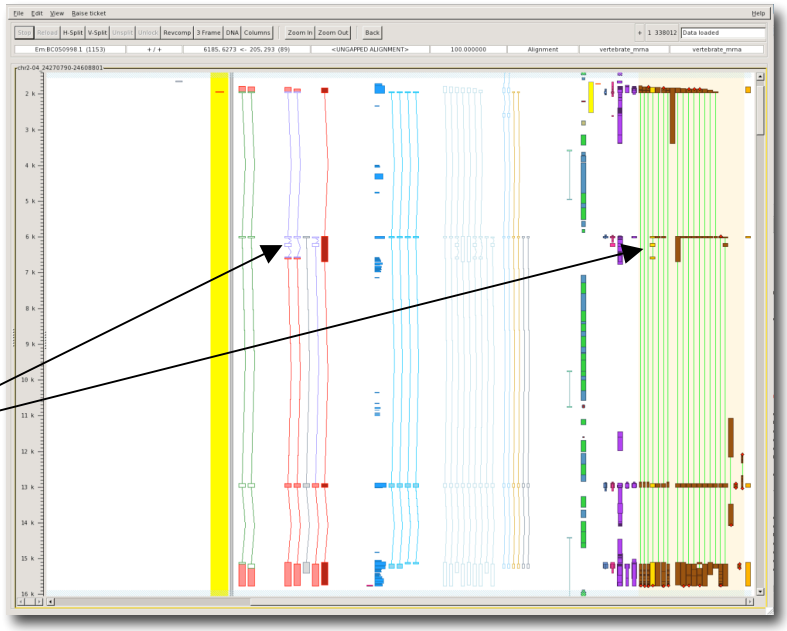

Recovered human chr2-04 clones 217..218
File SubSeq Clone Tools
Read Only
Stops Found in translation
Translation does not end with stop, and 'End not Found' is not set
No evidence attached
Find Clear



4) The evidence is attached automatically to the new gene object.



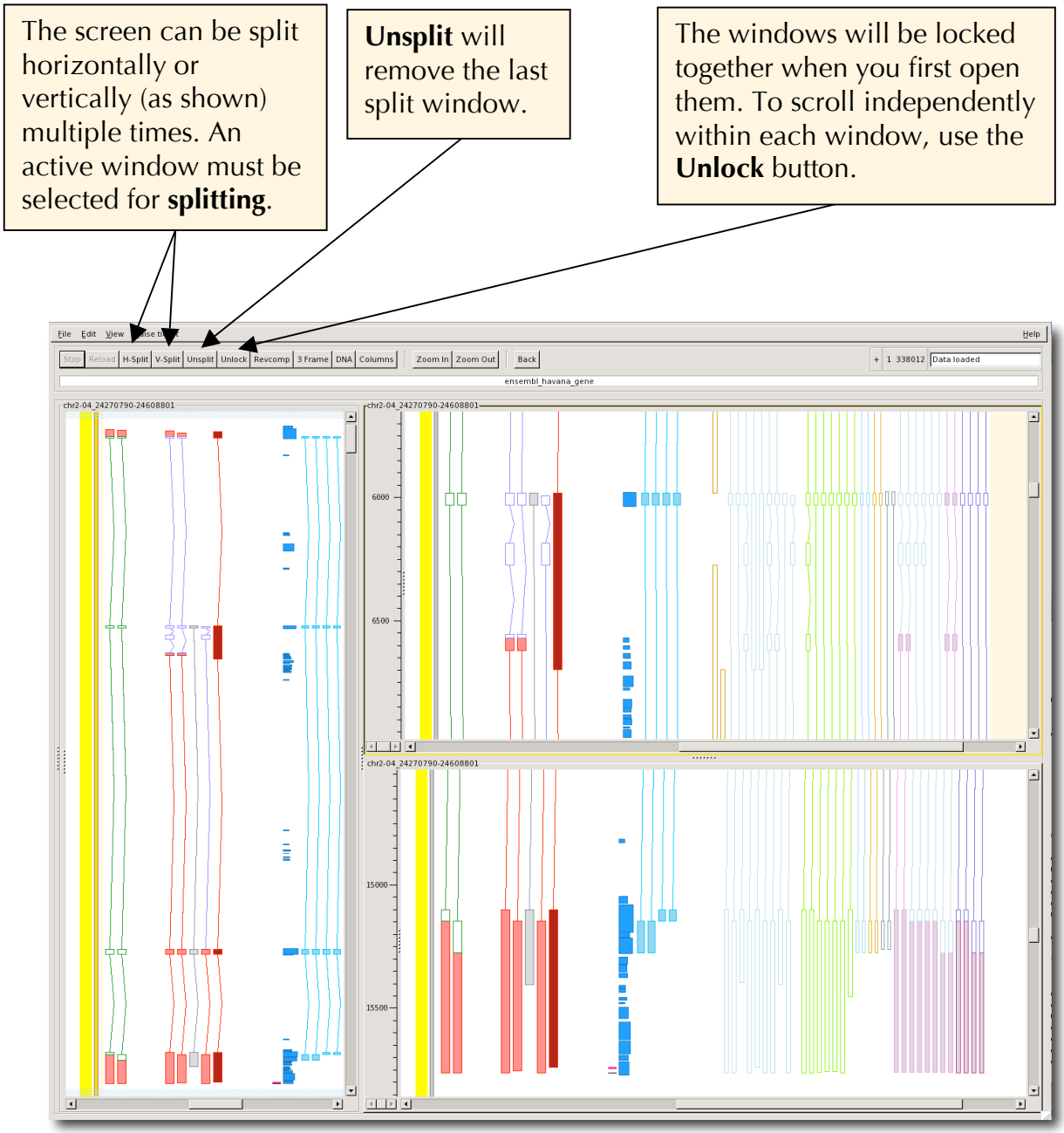
4) The new object will inherit its structure from the HSP. **However, you must always check the splice sites of your object in Blixem in case the alignment is incorrect.** Start/end coordinates (if a coding object) and transcript type are inherited from the parent, so these may not be relevant and may need to be changed. Note, that the new object is coloured red due to a number of errors. The checking software will not recognise evidence until the object is saved.



5) Once the errors have been removed, save the object to see it appear on Zmap (the evidence used has been highlighted).

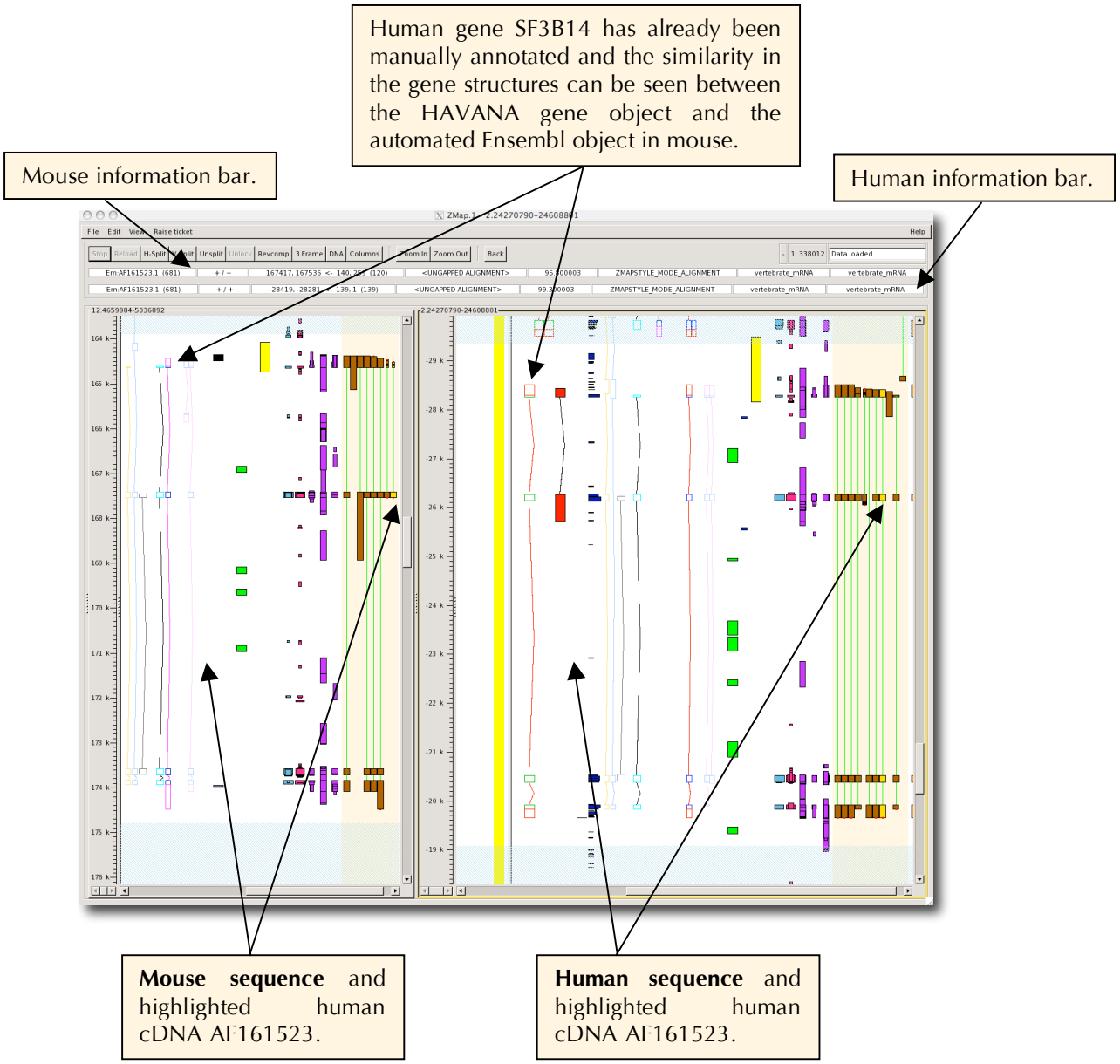
Splitting windows in Zmap

Use the **split** window function to effectively reduce the size of the window when looking at homologies. This is of particular use when you have to deal with very large introns because you can essentially reduce the introns to whatever size you wish, or when there are very many HSPs, because you can keep your gene object in view and static, but still scroll across the evidence.


The screen can be split horizontally or vertically (as shown) multiple times. An active window must be selected for **splitting**.
Unsplit will remove the last split window.
The windows will be locked together when you first open them. To scroll independently within each window, use the **Unlock** button.
File Edit View Help
Stop Reload H-Split V-Split Unsplit Unlock Revcomp 3 Frame DNA Columns Zoom In Zoom Out Back
1 338012 Data loaded
ensembl_havana_gene
chr2-04_24270790-24608801
6000
6500
15000
15500

Launching in a Zmap

This function allows you to open two or more sequences alongside each other (such as a human region and the syntenic region in mouse, or two haplotypes), so that simultaneous investigation can be carried out. To do this you will need to open both sets of clones in the same Otterlace session. To open both Zmap windows in one window as shown below, you need to select "Launch In A Zmap" option in one clone set. These clones will open to the left of the already open Otterlace session. This screen shot shows human gene SF3B14 and the syntenic region in mouse. The gene copy and paste function (referred to in the Otterlace section) is of much use here, saving time when building gene objects.


Human gene SF3B14 has already been manually annotated and the similarity in the gene structures can be seen between the HAVANA gene object and the automated Ensembl object in mouse.
Mouse information bar.
Human information bar.
Mouse sequence and highlighted human cDNA AF161523.
Human sequence and highlighted human cDNA AF161523.

Zmap keyboard and mouse shortcuts.

In general Zmap will be faster for zooming, bumping etc if you make good use of the built in short cuts. These can often avoid the need for Zmap to redraw large amounts of data that you may not even be interested in. For example, click once (highlight) on a feature and a carriage return will bring up evidence. Another example is to press T for translation.

All windows

Short Cut	**Action**
Cntl-W	close this window
Cntl-Q	quit ZMap

Zmap Window

Short Cut	**Action**
Control keys	
+ (or =), -	zoom in/out by 10%
Cntl + (or =), Cntl -	zoom in/out by 50%
up-arrow, down-arrow	scroll up/down slowly bit
Cntl up-arrow, Cntl down-arrow	scroll up/down more quickly
left-arrow, right-arrow	scroll left/right slowly
Cntl left-arrow, Cntl right-arrow	scroll left/right more quickly
page-up, page-down (Mac users should use fn and up/down arrow)	up/down by half a "page"
Cntl page-up, Cntl page-down	up/down by a whole "page"
Home, End (Mac users should use fn and left/rights arrows)	Go to far left or right
Cntl Home, Cntl End (Mac users will have to configure their keyboards for this)	Go to top or bottom
Delete, Shift Delete	Hide/Show selected features.
Enter	Show feature details for highlighted feature.
Shift up-arrow, Shift down-arrow	Jump from feature to feature within a column.
Shift left-arrow, Shift right-arrow	Jump from column to column.

Alpha-numeric keys

a	Blixem all sequences in column
A	Blixem only highlighted sequence in column

b	Bump/unbump current column within limits of mark if set, otherwise bump the whole column.
B	Bump/unBump current column within limits of the visible feature range.
c	compress/uncompress columns: hides columns that have no features in them either within the marked region or if there is no marked region within the range displayed on screen. Note that columns set to "Show" will not be hidden.
C	Compress/unCompress columns: hides all columns that have no features in them within the range displayed on screen regardless of any column, zoom, mark etc. settings.
h	Toggles highlighting (good for screen shots).
m	mark/unmark a range which spans whichever features or subparts of features are currently selected for zooming/smart bumping
M	Mark/unMark the whole feature corresponding to the currently selected subpart (e.g. the whole transcript of an exon or all HSPs of the same sequence as the highlighted one) for zooming/smart bumping
o or O	show menu Options for highlighted feature or column, use cursor keys to move through menu, press ESC to cancel menu.
r	reverse complement current view, complement is done for all windows of current view.
t or T	translate highlighted item, T hides Translation.
w or W	zoom out to show whole sequence
z	zoom to the extent of any selected features (e.g. exon/introns, HSPs etc) or any rubberbanded area if there was one.
Z	Zoom to whole transcript or all HSPs of a selected feature.

Zmap Mouse Usage

Left	**Middle**	**Right**
Single mouse button click		
highlight a feature or column Plus drag: draw a rectangle around an object for zoom	horizontal ruler with sequence position displayed, on button release centre on mouse position. Release mouse outside Zmap window to prevent re-centering.	show feature or column menu – for options such as pfetch, show feature DNA, show peptide, export peptide
Double mouse button click		
display details of selected feature. Double click on object to get edit window	same as single click	same as single click
Shift + mouse button click		
highlight a subpart of a feature (e.g. a single exon or alignment match)	same as single click	same as single click

OR multiple highlight		

Tips for a speedier Zmap

1. Specifically: zoom and mark within Zmap early on after launching. Either select a gene object and press 'z' to zoom OR select a rectangle to zoom in by dragging the left mouse button around it. Reverse complement now if necessary, then press 'm' to mark the region.

2. The quickest way to zoom out of Zmap again is to right mouse click on the 'zoom out' buttons at the top of zmap and choose one of the options (this is definitely much quicker that doing individual 'zoom outs' with the left mouse button). Likewise for 'zooming in' again (or use keyboard equivalents).

3. Bump within a marked region only. Bumping without marking is slow and removes the lines connecting Blast matches.

4. When you have finished working within a marked region, unbump the evidence you have been working on (e.g. ESTs) and unmark that region before you go on to select the next region to mark and bump – or you could miss visualising the evidence in the new region.

5. If you want to get rid of some white space try the compress 'c' function or alternatively toggle off some of the columns. Warning – this may hide features as well. If a column (e.g ESTs) is bumped and you want to lose it temporarily, it is quicker to turn the column off (when you turn it on again it will still be bumped when it re-appears) than unbump then rebump again later.

6. Jumping to genes/objects: If you expand the left hand 'scroll navigator' overview' you can jump directly to genes and objects by double-clicking on them.

Blixem User Manual

Written by Gemma Barson
(gb10@sanger.ac.uk)

Wellcome Trust Sanger Institute
17 January 2011

Blixem

This manual explains how to configure, run and use Blixem. Blixem is an interactive browser of pairwise matches displayed as multiple alignments. It is not strictly a multiple alignment tool, rather a 'one-to-many' alignment. It is used to check the alignments of nucleotide and amino acid sequences against a reference sequence.

Blixem is maintained by the Wellcome Trust Sanger Institute and is available as part of the SeqTools package. The software can be downloaded from the Sanger Institute's website: http://www.sanger.ac.uk.

An aside about the name "Blixem"
"BLIXEM" was originally an acronym for "BLast matches In an X-windows Embedded Multiple alignment", although this is a bit of a misnomer now because Blixem can handle any kind of alignment, not just BLAST matches. We have dropped the acronym, and the capital letters, so the correct name is just "Blixem".

Getting Started

Running Blixem

As a minimum, Blixem takes the following required arguments:

```
blixem –-display-mode N|P <features_file>
```

Where `<features_file>` is the path name of a GFF version 3 file containing the alignments and any other features.

The '`--display-mode`' argument is the only mandatory option. It defines the display mode: '`N`' for nucleotide or '`P`' for protein. .

Run '`blixem`' without any arguments to see further usage information.

Input files

Blixem takes one or two files as input: a mandatory GFF version 3 file containing the features and, optionally, a separate file containing the reference sequence in FASTA format.

```
blixem –m N|P [<reference_sequence_file>] <features_file>
```

If the reference sequence file is not provided, the reference sequence must be supplied in FASTA format at the end of the GFF file, following a comment line that reads '`##FASTA`'.

Note that the reference sequence must always be a nucleotide sequence and match sequences must be the correct type for the mode, i.e. nucleotide sequences for nucleotide mode or protein sequences for protein mode.

GFF file

Blixem uses the GFF version 3 file format. In this section we give a very brief description of this file format; see http://www.sequenceontology.org/gff3.shtml for a full description.

The GFF file should start with the following two comment lines. (Additional comments can be included but may be ignored.)

```
##gff-version 3
##sequence-region chr4-04_210623-364887 44144 154265
```

Each subsequent line defines a feature. A feature line must have the following 8 tab-separated columns:

```
reference_sequence_name       source  type    start   end     score   strand  phase
```

An optional 9th column defines any tags (separated by semi-colons). Blixem supports the following GFF tags. (Additional tags can be supplied but may be ignored.)

Target (required for alignments)
Gap (required for gapped alignments)
ID (required for parent features)
Name (required for transcripts and SNPs)
Parent (required for child features)

In addition, Blixem supports the following custom tags.

percentId (only applicable to alignments; populates the %ID column)
sequence (only applicable to alignments; supplies the sequence data)
variant_sequence (only applicable to variations; supplies the variation data)
url (only used by variations; GFF3 special characters must be escaped)

Transcripts

Note that exons should have a Parent transcript defined, and the Name tag should be set in the parent rather than the child exons. Note that Blixem *will* recognise exons that do not have a Parent tag if they have a Name tag instead, but they may not get grouped correctly with other exons from the same transcript.

Typically, one defines the parent transcript, the exons, and the CDS regions; Blixem will then calculate the missing components (in this case, the UTR regions and the introns). Blixem will recognise other combinations of inputs, and will always calculate the missing components as long as enough information is provided.

Variations

SNPs, insertions and deletions are supported, as well as combined variations. One may use the generic '`sequence_alteration`' type for these but it is good practice to use more specific types such as '`SNP`' or '`deletion`' where applicable.

Sample GFF file

A sample GFF file may look like this ('…' denotes that text has been omitted).

```
##gff-version 3
##sequence-region chr4-04_210623-364887 44144 154265
chr4-04_210623-364887  EST_Human      nucleotide_match       79195   79311   95.000000
       -       .       Target=DA692754.1 287 403 +;percentID=90.6;sequence=GATCTGGC...
chr4-04_210623-364887  EST_Human      nucleotide_match       79195   79323   121.000000
       +       .       Target=AI095103.1 326 454 +;percentID=96.9;sequence=TTTAAATT...
chr4-04_210623-364887  ensembl_variation      deletion       80798   80799   .       +
       .
       Name=rs60725655;url=http%3A%2F%2Fwww.ensembl.org%2FHomo_sapiens%2FVariation%2FSumm
ary%3Fv%3Drs60725655;variant_sequence=AA/-;
chr4-04_210623-364887  ensembl_variation      sequence_alteration    80799   80799   .
       +       .
       Name=rs57681246;url=http%3A%2F%2Fwww.ensembl.org%2FHomo_sapiens%2FVariation%2FSumm
ary%3Fv%3Drs57681246;variant_sequence=A/-/C;
chr4-04_210623-364887  ensembl_variation      SNP     81040   81040   .       +       .
       Name=rs2352935;url=http%3A%2F%2Fwww.ensembl.org%2FHomo_sapiens%2FVariation%2FSumma
ry%3Fv%3Drs2352935;variant_sequence=T/C;
```

```
chr4-04_210623-364887  ensembl_variation     insertion      82229   82230   .       +
       .
       Name=rs35105663;url=http%3A%2F%2Fwww.ensembl.org%2FHomo_sapiens%2FVariation%2FSumm
ary%3Fv%3Drs35105663;variant_sequence=-/G;
chr4-04_210623-364887  Augustus       mRNA    119534  119941  .       -       .
       ID=transcript21;Name=AUGUSTUS00000051712
chr4-04_210623-364887  Augustus       exon    119534  119941  .       -       .
       Parent=transcript21
chr4-04_210623-364887  Augustus       CDS     119534  119941  .       -       0
       Parent=transcript21
```

FASTA file

A FASTA file has a header line that starts with '>' and contains the sequence name. The next line contains the start of the sequence data. The sequence data can be on a single line or separated by newlines; it is usually separated by newlines every 50 characters to aid readability.

```
>chr4-04_210623-364887
tcttgtttctgtaggagaggccatctccatcagctataaccaaaaaaaaa
acaaaaaactcctctttttgacaagtttgtaaagcctgtccatctgggtc
tataataatcctccaggccctatgccactcctctttattcagccagttca
...
```

Combined GFF and FASTA file

```
##gff-version 3
##sequence-region chr4-04_210623-364887 44144 154265
chr4-04_210623-364887  EST_Human      nucleotide_match       79195   79311   95.000000
       -       .       Target=DA692754.1 287 403 +;percentID=90.6
chr4-04_210623-364887  EST_Human      nucleotide_match       79195   79323   121.000000
       +       .       Target=AI095103.1 326 454 +;percentID=96.9
...
##FASTA
>chr4-04_210623-364887
tcttgtttctgtaggagaggccatctccatcagctataaccaaaaaaaaa
acaaaaaactcctctttttgacaagtttgtaaagcctgtccatctgggtc
tataataatcctccaggccctatgccactcctctttattcagccagttca
...
```

Configuration file

Note that if the sequence data for the match sequences is not supplied via the 'sequence' tag in the GFF file then Blixem will try to fetch the data from a server using a program called 'pfetch'. Currently this is only supported for internal users at the Sanger Institute. Details of the server are supplied via a .ini-style configuration file using the '-c' argument.

The Blixem Window

The Blixem window consists of two main sections: an overview section called the "big picture", and a detail section showing the actual sequence data. These sections are separated by a splitter bar, so you can maximise the space for the area you are interested in. You can also hide sections of the window using the 'View' menu.

Blixem can show sequences in nucleotide or protein mode.

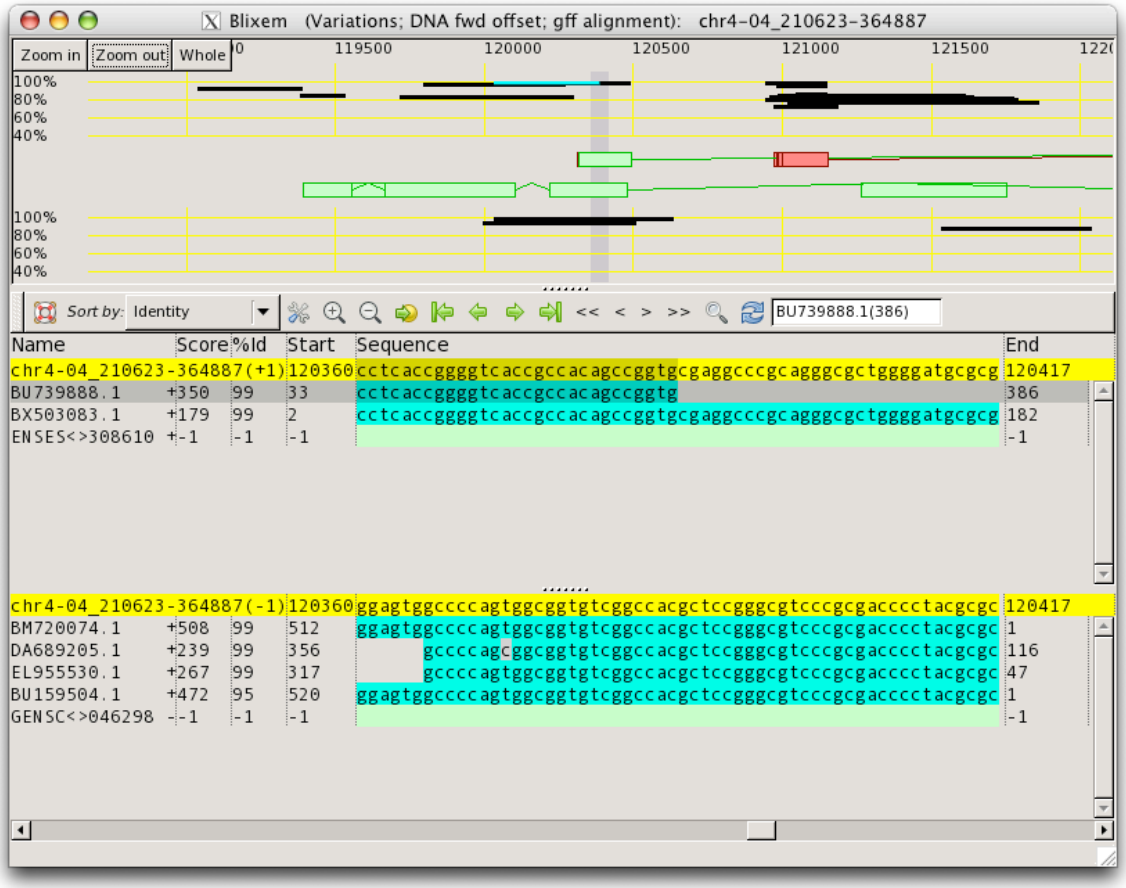


Figure 1: Nucleotide mode. There are two panes in the detail-view, one for each strand. The active strand is shown at the top. The active strand can be changed by hitting the 'Toggle' button or the 't' shortcut key.

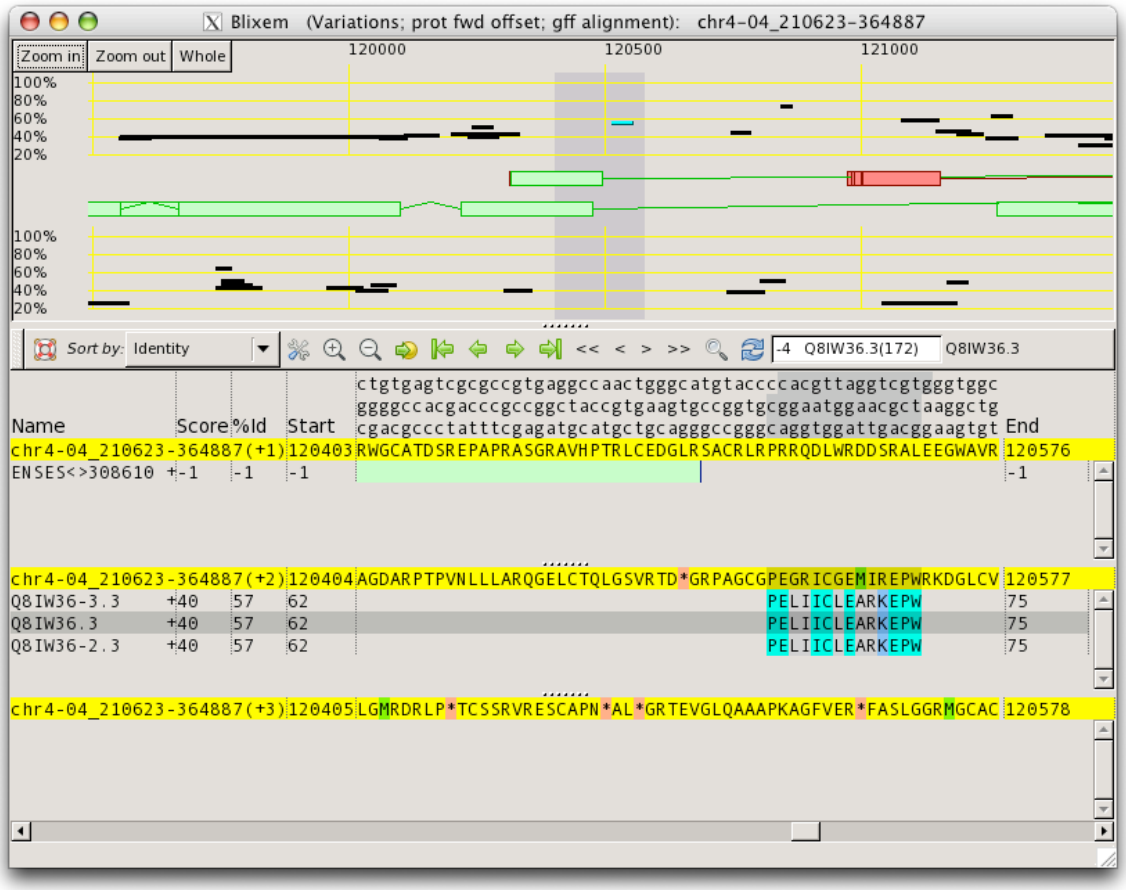


Figure 2: Protein mode. There are three panes in the detail-view; one for each reading frame of the active strand. The other strand can be activated by hitting the 'Toggle' button or the 't' shortcut key.

Active Strand

The "active" reference sequence strand in Blixem controls the orientation of the display – coordinates are shown increasing from left-to-right for the forward strand and decreasing for the reverse strand. The active strand is always shown at the top – i.e. the top grid and top transcript view in the big picture and the top pane in the detail view.

In protein mode, only the active strand is shown in the detail view. One must toggle the strand to view the other strand.

Toggle which strand is active by:

- pressing the 'Toggle' button on the toolbar; or
- pressing the 't' key.

By default, Blixem assumes that the reference sequence passed to it is the forward strand, unless otherwise specified by the '--reverse-strand' command line argument.

Big Picture

The 'Big Picture' section shows an overview of the reference sequence. The reference sequence coordinates are shown along the top. You can zoom in to view a shorter range by using the 'Zoom in' button at the top left of the screen. Use 'Zoom out' or 'Whole' to zoom out – 'Whole' zooms out to view the full length of the reference sequence.

The big picture consists of two grids showing the alignments for each strand, and two sections between these grids showing the transcripts for each strand. The grids have a scale on the left-hand side showing the percent-ID, and alignments are plotted against this scale. The scale and extents of the grids can both be edited - see the section in the Settings dialog.

The active strand alignments and transcripts are shown at the top and the other strand at the bottom. The direction of the coordinates is determined by the active strand. The active strand can be toggled using the 't' shortcut key or the 'Toggle strand' button on the toolbar.

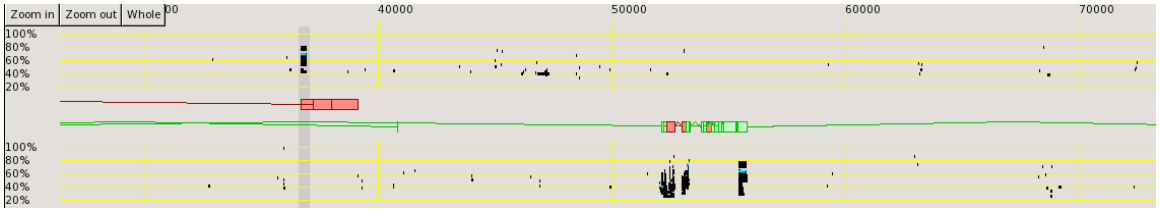


Figure 3: The Big Picture section

Bumping the transcript view

By default, exons and introns for the same strand are drawn overlapping each other. They can be expanded (or 'bumped') by pressing the 'b' shortcut key or by enabling the relevant option in the View dialog (see).

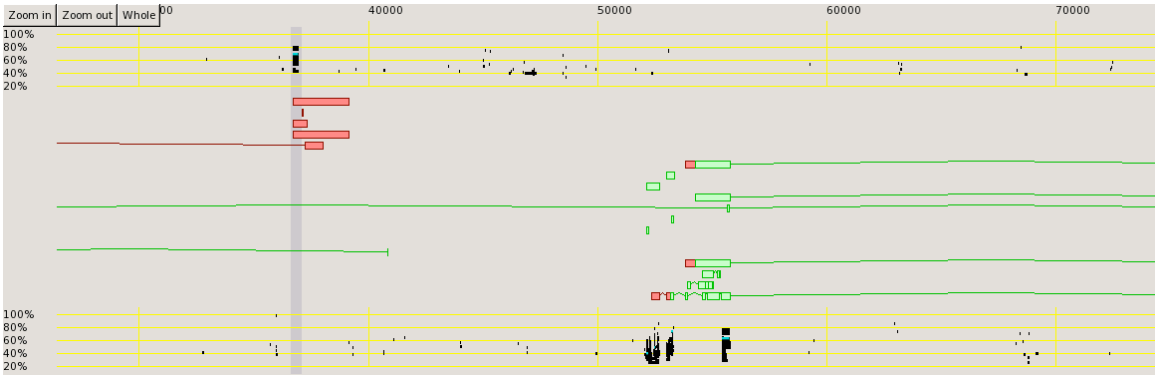


Figure 4: Expanded transcript view

Detail View

The 'Detail View' shows the actual sequence data for the match sequences. Match sequences are lined up underneath the relevant section of reference sequence, and individual bases are highlighted in different colours to indicate how well they match.

Match colours

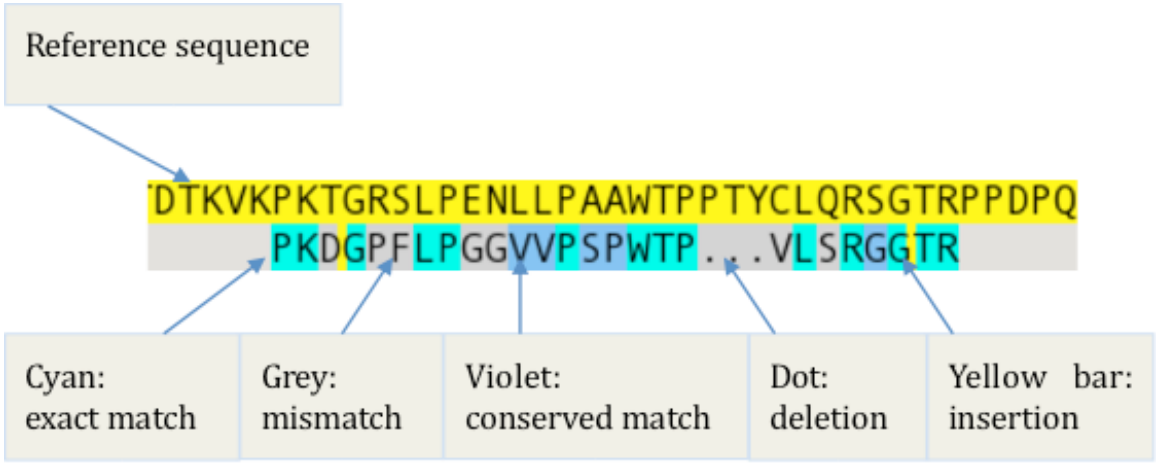


Figure 5: Alignment colour key

Alignment lists

There are separate lists of alignments for each strand and reading frame of the reference sequence. Each list has a yellow header bar containing the reference sequence. At the left, the yellow bar shows the reference sequence name and which strand/frame it is, e.g. (+1) means forward strand, reading frame 1; (-2) means reverse strand, reading frame 2.

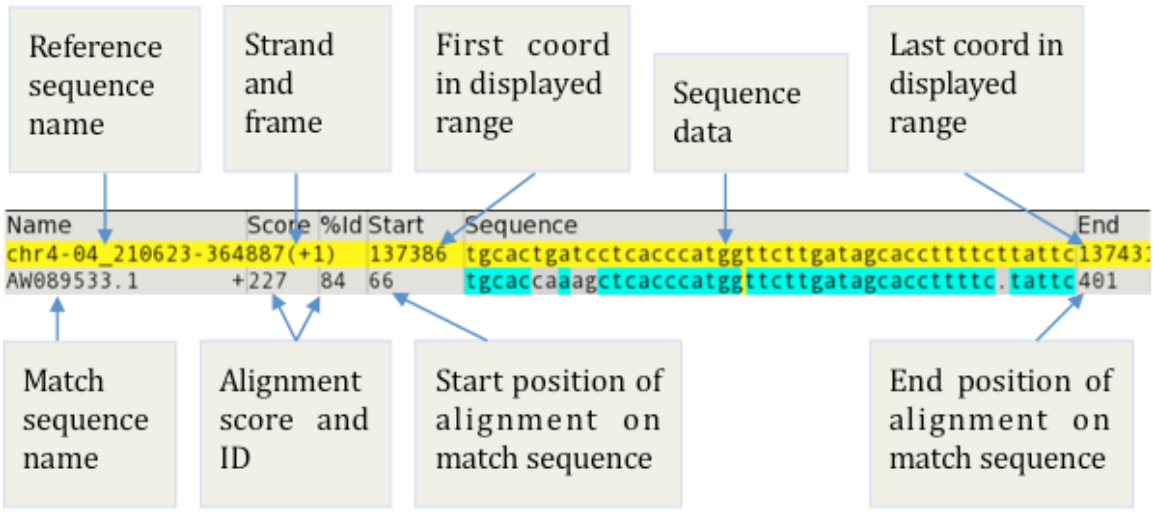

Reference sequence name
Strand and frame
First coord in displayed range
Sequence data
Last coord in displayed range
Name Score %Id Start Sequence End
chr4-04_210623-364887(+1) 137386 tgcactgatcctcacccatggttcttgatagcacctttttcttattc 13743
AW089533.1 +227 84 66 tgcaccaaagctcacccatggttcttgatagcacctttttc.tattc 401
Match sequence name
Alignment score and ID
Start position of alignment on match sequence
End position of alignment on match sequence

Figure 6: Alignment list details

Nucleotide mode

There are two sections to the detail view in nucleotide mode: one for each strand. The active strand is shown at the top and defines the coordinate direction (increasing if the forward strand is active, decreasing if the reverse is active).

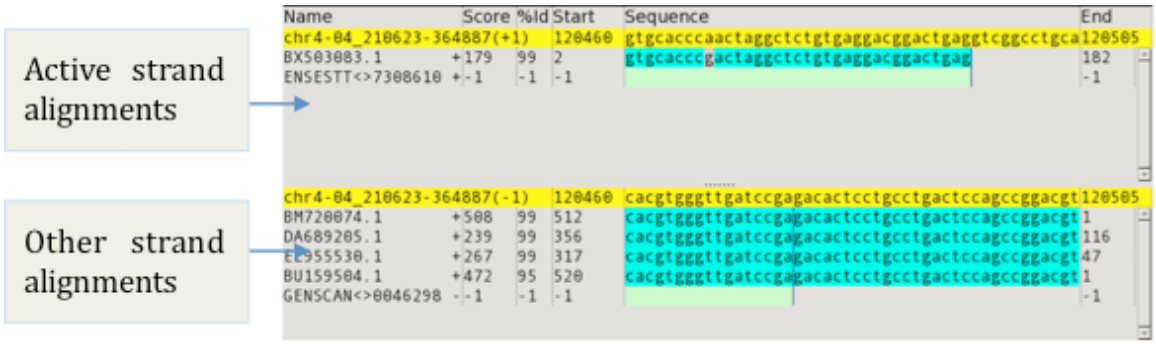

Active strand alignments
Other strand alignments
Name Score %Id Start Sequence End
chr4-04_210623-364887(+1) 120460 gtgcacccaactaggctctgtgaggacggactgaggtcggcctgca 120505
BX503083.1 +179 99 2 gtgcacccgactaggctctgtgaggacggactgag 182
ENSESTT<>7308610 +-1 -1 -1 -1
chr4-04_210623-364887(-1) 120460 cacgtgggttgatccgagacactcctgcctgactccagccggacgt 120505
BM720074.1 +508 99 512 cacgtgggttgatccgagacactcctgcctgactccagccggacgt 1
DA689205.1 +239 99 356 cacgtgggttgatccgagacactcctgcctgactccagccggacgt 116
EE955530.1 +267 99 317 cacgtgggttgatccgagacactcctgcctgactccagccggacgt 47
BU159504.1 +472 95 520 cacgtgggttgatccgagacactcctgcctgactccagccggacgt 1
GENSCAN<>0046298 --1 -1 -1 -1

Figure 7: Alignment lists: nucleotide mode

Protein mode

There are three sections in the detail view in protein mode: one for each of the three reading frames for the active strand. Only the active strand is shown. To view the other strand, toggle the display using the 'Toggle strand' button or the 't' shortcut key.

In protein mode, the yellow header bars show the translated reference sequence for that reading frame. STOP and MET codons in the reference sequence are highlighted in red and green. There is also an additional header section at the top showing the nucleotide sequence.

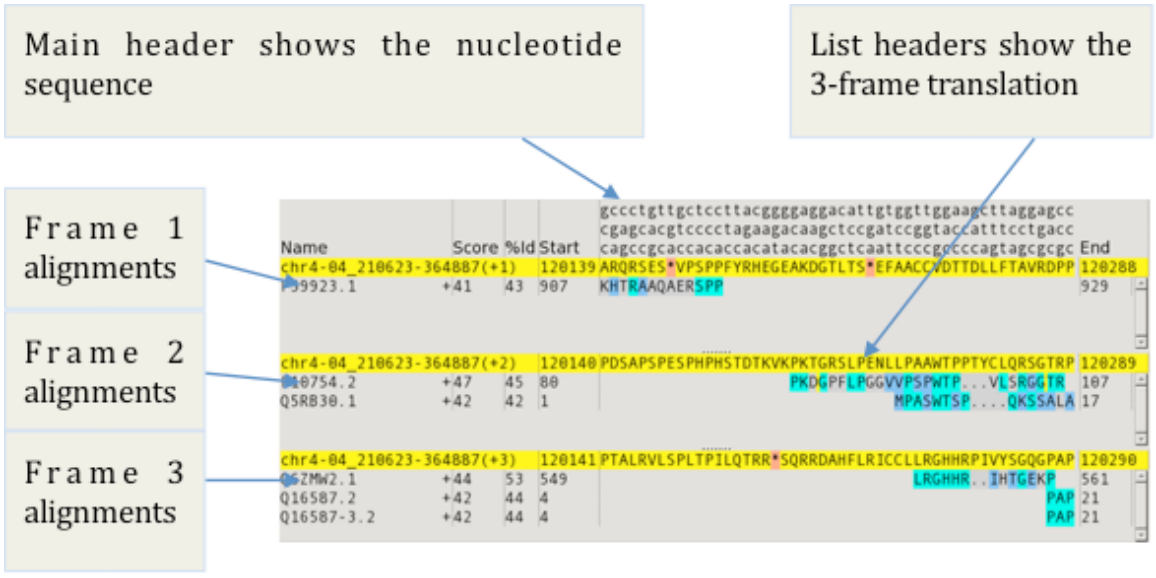

Main header shows the nucleotide sequence
List headers show the 3-frame translation
Frame 1 alignments
Frame 2 alignments
Frame 3 alignments
Name Score %Id Start End

Figure 8: Alignment lists: protein mode

In the nucleotide-sequence header, codons are read from top-to-bottom and then left-to-right, starting at row 1 for frame 1, row 2 for frame 2 etc. Middle-clicking on a coordinate will highlight the three nucleotides for the selected codon and the currently-active reading frame (by default, frame 1). Left-clicking in an alignment list sets the active reading frame.

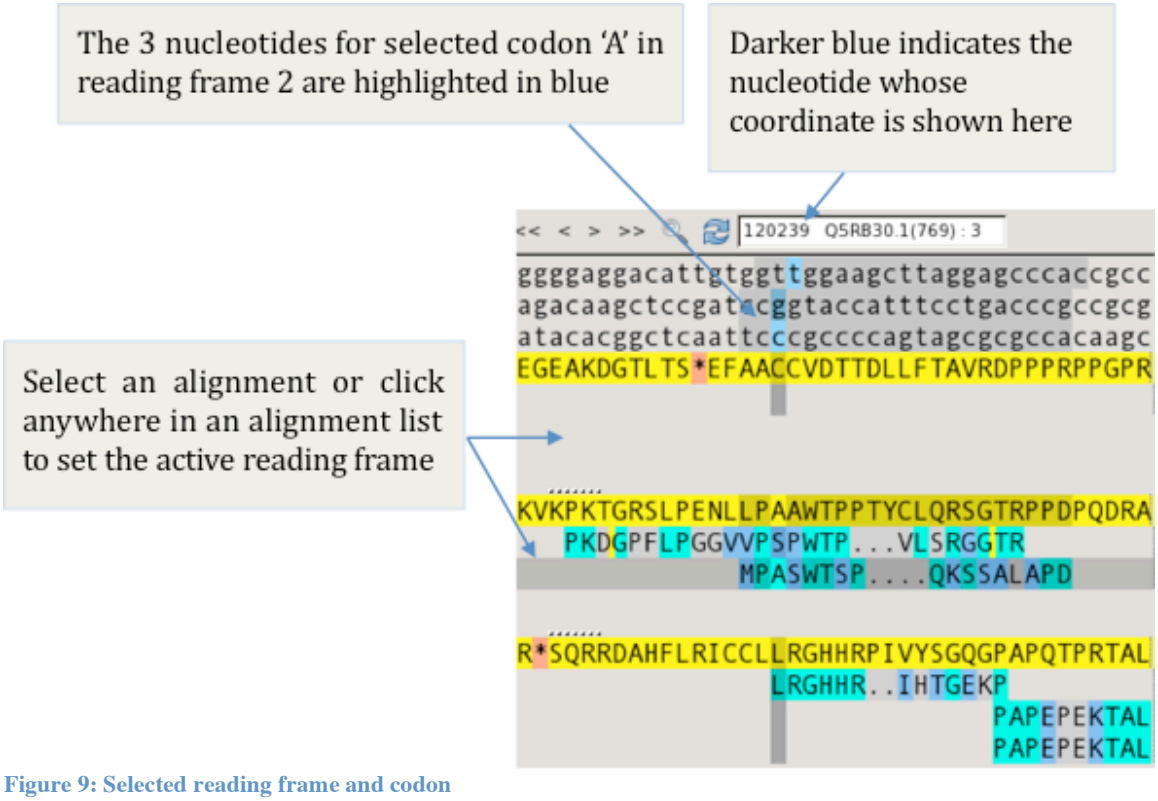

The 3 nucleotides for selected codon 'A' in reading frame 2 are highlighted in blue
Darker blue indicates the nucleotide whose coordinate is shown here
Select an alignment or click anywhere in an alignment list to set the active reading frame
120239 Q5RB30.1(769) : 3

Figure 9: Selected reading frame and codon

The toolbar

The detail-view toolbar contains the following functions. Note that the Help and Settings buttons are included in the detail-view toolbar even though they apply to Blixem as a whole.

Figure 5: Detail-view toolbar

Help:	Show help about how to use Blixem
Sort-by:	Select which column to sort the match sequences by
Settings:	Show the Settings dialog.
Zoom in:	Increase the font size in the detail-view
Zoom out:	Decrease the font size in the detail-view
Go to:	Go to a particular coordinate
First match:	Go to the first coordinate of the first alignment[1]
Previous match:	Go to the start of the current alignment or the end of the previous alignment[1]
Next match:	Go to the end of the current alignment or the start of the next alignment[1]
Last match:	Go to the end of the last alignment[1]
Back one page:	Scroll the detail-view range to the left by one page
Back one index:	Scroll the detail-view range to the left by one base
Forward one index:	Scroll the detail-view range to the right by one base
Forward one page:	Scroll the detail-view range to the right by one page
Find:	Scrolls to the start of the first alignment from that

[1] Acts only on selected sequences, if there is currently a selection; if no sequences are currently selected, then this operation acts on all sequences.

sequence if any are found.

Toggle strand: Toggle which strand is the active strand

Feedback box

The feedback box contains information about the currently selected sequence and/or coordinate, if either is selected. Click on a row in the detail-view to select a sequence. Middle-click on a base in the detail-view to select that coordinate. Text in the feedback box can be selected and copied.

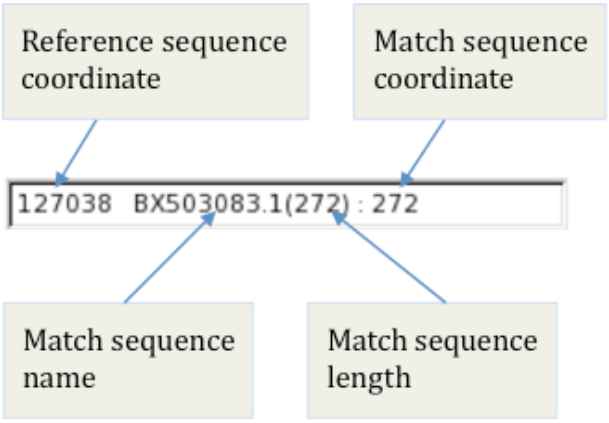


Figure 11: Feedback box

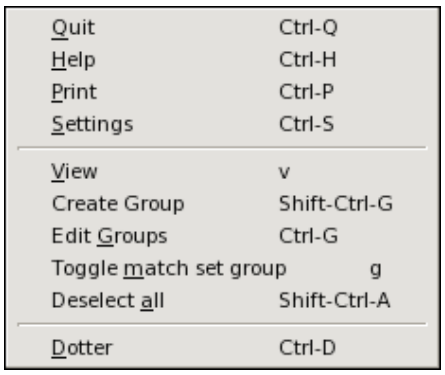
Moused-over item feedback area

The area to the right of the toolbar contains information about the currently moused-over item (e.g. a match sequence in the alignment list or a variation in the variations track). For a match sequence, this information includes the sequence name and optional data such as organism and tissue type that can be parsed from EMBL files (currently only available to authorised users). To load optional data, see the Settings dialog. Note that the optional data may be incomplete due to the inconsistent information available from the EMBL files.

Error! Bookmark not defined.

The main menu

Right-click anywhere in the Blixem window to pop up the main menu.



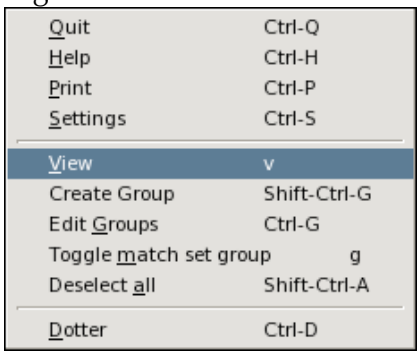
The options are:

Quit	*Ctrl-Q*	Close Blixem and any spawned processes
Help	*Ctrl-H*	Display the user help
Print	*Ctrl-P*	Printing options
Settings	*Ctrl-S*	Edit settings
View	*v*	Show/hide parts of the display
Create Group	*Shift-Ctrl-G*	Create a group of sequences
Edit Groups	*Ctrl-G*	Edit properties for groups
Deselect all	*Shift-Ctrl-A*	Deselect all sequences
Dotter	*Ctrl-D*	Run Dotter on the currently selected sequence

Hiding sections of the window

Use to 'View' dialog to show/hide sections of the window.

1. Right-click and select the View option, or hit the 'v' shortcut key.



2. Toggle check marks on or off to show/hide sections.

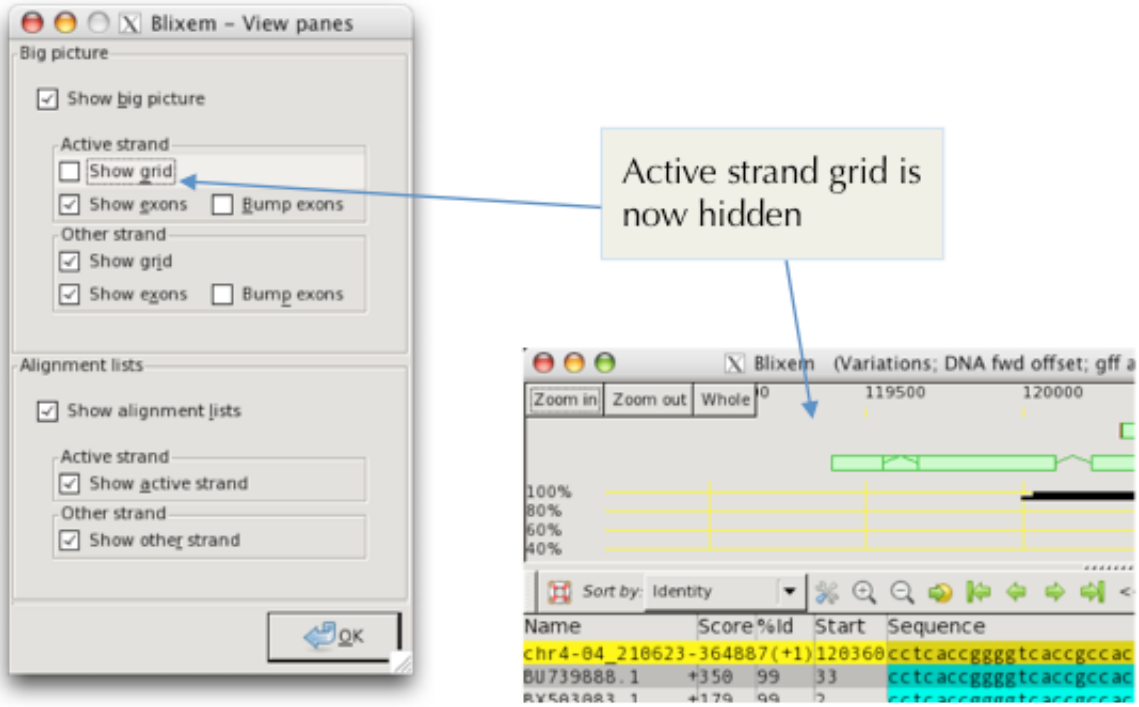


Figure 13: The View dialog

Alternatively, use the following keyboard shortcuts to toggle visibility of a component:

1	Hide top pane in detail view
2	Hide second pane in detail view
3	Hide third pane in detail view (protein mode only)
Ctrl-1	Hide top grid in big picture (active strand)
Ctrl-2	Hide bottom grid in big picture (other strand)
Shift-Ctrl-1	Hide top exon view (active strand)
Shift-Ctrl-2	Hide bottom exon view (other strand)

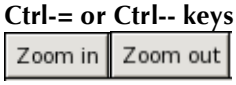
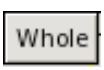
Operation

Navigation

Scrolling

Middle-click/drag in big picture	Select a region to jump to.
Middle-click/drag in detail view	Select and centre on a base.
Horizontal scrollbar	Scroll the detail-view range.
Vertical scrollbars	Scroll up/down an alignment list.
Horizontal mouse-wheel	Scroll the detail-view range (if your mouse has a horizontal scroll-wheel).
Vertical mouse-wheel	Scroll up/down the currently moused-over alignment list
Ctrl-left **Ctrl-right**	Scroll to the start/end of the previous/next match (limited to currently-selected sequences, if any are selected; includes all sequences otherwise).
Home **End**	Scroll to the start/end of the display.
Ctrl-Home **Ctrl-End**	Scroll to the start/end of the currently-selected alignments (or to the first/last alignment if none are selected).
',' (comma) **'.' (full-stop)**	Scroll the detail-view range one nucleotide to the left/right.
Ctrl-, **Ctrl-.**	Scroll the detail-view range one page to the left/right.
Go-to button or 'p' key	Scroll to a specific coordinate position.

Zooming

= - keys and	Zoom in/out of the detail-view
Ctrl-= or Ctrl-- keys and 	Zoom in/out of the big-picture
Shift-Ctrl-- and 	Zoom the big picture out to view the full length of the reference sequence.

Selections

Selecting sequences

- You can select a sequence by clicking on its row in the alignment list. Selected sequences are highlighted in cyan in the big picture.
- You can select a sequence by clicking on it in the big picture.
- The name of the sequence you selected is displayed in the feedback box on the toolbar. If there are multiple alignments for the same sequence, all of them will be selected.
- You can select multiple sequences by holding down the Ctrl or Shift keys while selecting rows.
    - You can deselect a single sequence by Ctrl-clicking on its row.
- You can deselect all sequences by right-clicking and selecting 'Deselect all', or with the Shift-Ctrl-A keyboard shortcut.
- You can move the selection up/down a row using the up/down arrow keys.

Selecting coordinates

- You can select a nucleotide/peptide by middle-clicking on it in the detail view. This selects the entire column at that index, and the coordinate number on the reference sequence is shown in the feedback box. (The coordinate on the match sequence is also shown if a match sequence is selected.)
- By default the display will centre on the selected base when you middle click. To select a base without scrolling, hold down Ctrl when you middle click.
- For protein matches, when a peptide is selected, the three nucleotides for that peptide (for the active reading frame) are highlighted in the header in blue. (The active reading frame is whichever alignment list currently has the focus - click in a different list to change the reading frame.) Darker blue highlighting indicates the specific nucleotide that is currently selected (i.e. whose coordinate is displayed in the feedback box).

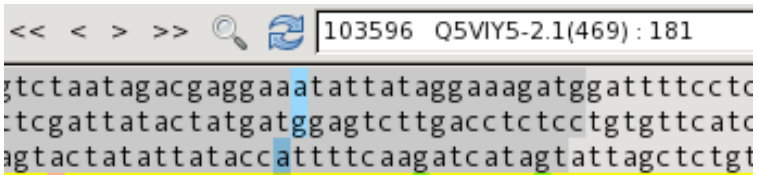


Figure 6: The 3 nucleotides for the currently-selected amino acid in reading-frame 3. Selected nucleotide 103596 is shaded in darker blue.

- You can move the selection to the previous/next index using the left and right arrow keys.
- In protein mode, you can move the selected nucleotide by a single base (rather than an entire codon) holding Shift while using the left and right arrow keys.

- You can move the selection to the start/end of the previous/next match by holding Ctrl while using the left and right arrow keys (limited to just the selected sequences if any are selected; includes all sequences otherwise).

Finding sequences

The Find dialog allows the user to search for sequences by name. Press the Find button on the toolbar or hit the 'Ctrl-F' shortcut key to open the Find dialog.

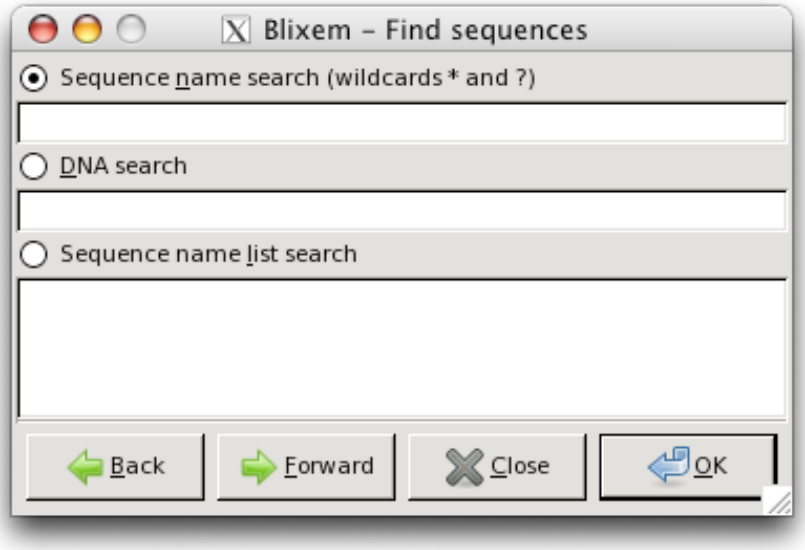


Figure 7: Find dialog

There are three search modes:

- Sequence name search: Search for match sequences by name. The wild-card '*' means any number (or zero) of any character and '?' means 1 character (which can be any character). Any sequences whose names match the search string will be selected and the display will scroll to the start of the selection.
- DNA search: This searches for a given sub-sequence of nucleotides in the reference sequence. If the sub-sequence is found, the display will scroll to the start of the sub-sequence and the first base in the sub-sequence will be selected.
- Sequence name list search: the same as 'Sequence name search', but for multiple sequences. Each sequence names should be on a separate line.

Enter your search text in the appropriate box and click the OK button to perform the search. By default, Blixem will start the search at the beginning of the reference sequence range. To start the search from the current position, click the Forward or Back button instead of OK. This will start searching from the currently-selected base, if there is one selected; if not, it will start from the beginning of the current

detail-view display range when searching forwards or from the end of the display range if searching backwards.

Repeat a Find

After clicking OK on the Find dialog, press F3 to repeat the search in a forwards direction or Shift-F3 to repeat in a backwards direction. Alternatively, if you had selected the Forward or Back button in the Find dialog then click the Forward or Back buttons again to jump to the next result in that direction.

Copy and paste

- When sequence(s) are selected, their names are copied to the *selection buffer* and can be pasted to another program by middle-clicking in that program.
- Sequence names can be pasted from the selection buffer into Blixem by hitting the 'f' keyboard shortcut. If the selection buffer contains valid sequence names, those sequences will be selected and the display will jump to the start of the selection.
- Sequence names can also be pasted from the selection buffer into text boxes in dialog boxes such as the Groups dialog or Find dialog.
- To copy sequence name(s) to the *default clipboard*, select the sequence(s) and hit Ctrl-C. Sequence names can then be pasted into other applications using Ctrl-V.
- The default clipboard can be pasted into Blixem using Ctrl-V. If the clipboard contains valid sequence names, those sequences will be selected and the display will jump to the start of the selection.
- Note that text from the feedback box and some text labels (e.g. the reference sequence start/end coords) can be copied to the selection buffer by selecting the required text with the mouse (or copied to the default clipboard by selecting it and then hitting 'Ctrl-C').
- Text can be pasted from the default clipboard into text entry boxes on dialogs such as the Groups or Find dialog by using Ctrl-V.

Sorting alignments

- Alignments can be sorted by selecting the column you wish to sort by from the drop-down box on the toolbar.

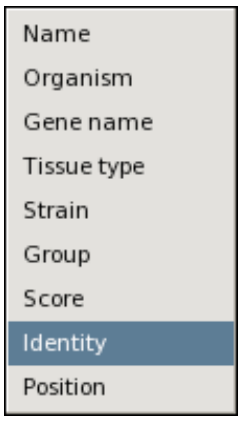


Figure 8: Sort-by list

- The default sort order may be ascending or descending depending on what makes most sense for the selected column: e.g. sorting by position is ascending by default but sorting by score or ID is descending.
- To get the inverse of the default sort order, select the 'Invert sort order' option in the Settings dialog.
- Alignments can also be sorted by group. Alignments that are part of a group will then be listed first (before any that are not in a group), and ordered according to the group's order number. See the Groups section for more details.

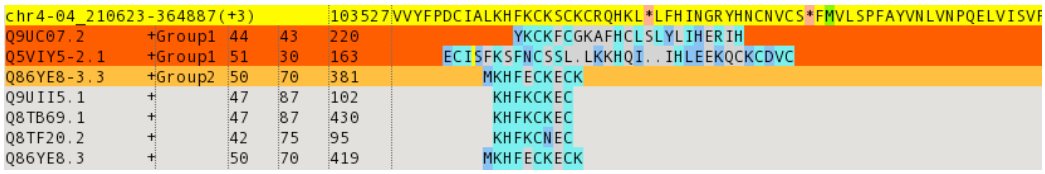

chr4-04_210623-364887(+3) 103527VVYFPDCIALKHFKCKSCKCRQHKL*LFHINGRYHNCNVCS*FMVLSPFAYVNLVNPQELVISVF
Q9UC07.2 +Group1 44 43 220 YKCKFCGKAFHCLSLYLIHERIH
Q5VIY5-2.1 +Group1 51 30 163 ECISFKSFNCSSL.LKKHQI..IHLEEKQCKCDVC
Q86YE8-3.3 +Group2 50 70 381 MKHFECKECK
Q9UII5.1 + 47 87 102 KHFKCKEC
Q8TB69.1 + 47 87 430 KHFKCKEC
Q8TF20.2 + 42 75 95 KHFKCNEC
Q86YE8.3 + 50 70 419 MKHFECKECK

Figure 9: Alignment list sorted by group

Fetching sequences

Currently only available to authorised users at the Sanger Institute.

- Double-click a row to fetch a match sequence's EMBL file.

Grouping sequences

Alignments can be grouped together so that they can be sorted/highlighted/hidden etc.

Creating a group from a selection:

- Select the sequences you wish to include in the group by left-clicking their rows in the detail view. Multiple rows can be selected by holding the Ctrl or Shift keys while clicking.
- Right-click and select 'Create Group', or use the Shift-Ctrl-G shortcut key. (Note that Ctrl-G will also shortcut to here if no groups currently exist.)

- Ensure that the 'From selection' radio button is selected, and click 'OK' or 'Apply'. If you click 'Apply', you will be shown the group you just created so that you can edit it. If you click 'OK' the group will be created with the default properties.

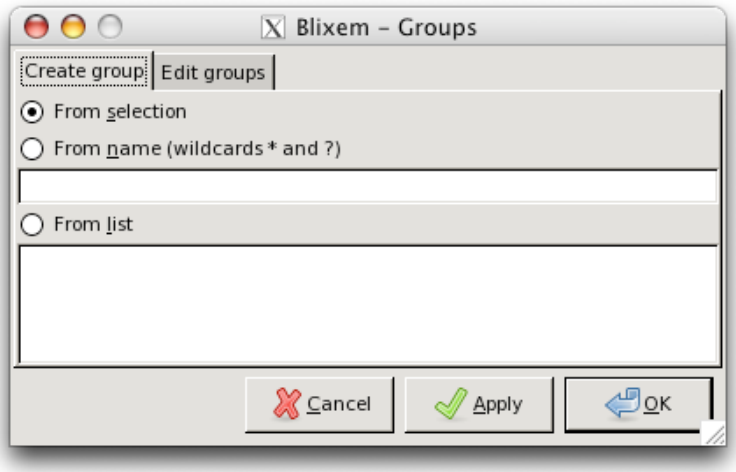


Figure 10: Groups dialog: create group

Creating a group from a sequence name:

- Right-click and select 'Create Group', or use the Shift-Ctrl-G shortcut key. (Or Ctrl-G if no groups currently exist.)
- Select the 'From name' radio button and enter the name of the sequence in the box below. You may use the following wildcards to search for sequences: '*' for any number of characters; '?' for a single character.
- Click 'OK'.

Creating a group from sequence name(s):

- Right-click and select 'Create Group', or use the Shift-Ctrl-G shortcut key. (Or Ctrl-G if no groups currently exist.)
- Select the 'From name(s)' radio button.
- Enter the sequence name(s) in the text box.
- You may use the following wild-cards in a sequence name: '*' for any number of characters; '?' for a single character.
- You may search for multiple sequence names by separating them with the following delimiters: newline, comma or semi-colon.
- You may paste sequence names directly from another compatible program (e.g. ZMap): click on the feature in ZMap and then middle-click in the text box on the Groups dialog. (Grouping in Blixem works on the sequence name alone, so the feature coords output by ZMap will be ignored.)
- Click 'OK'.

Creating a temporary 'match-set' group from the current selection:

- You can quickly create a group from a current selection (e.g. selected features in ZMap or just the current selection in Blixem) using the 'Toggle match set' option.
- To create a match-set group, select the required items and then select 'Toggle match set' from the right-click menu in Blixem, or hit the 'g' shortcut key.
- To clear the match-set group, choose the 'Toggle match set' option again, or hit the 'g' shortcut key again.
- While it is enabled (i.e. toggled on), the match-set group can be edited like any other group, via the 'Edit Groups' dialog. Any settings you change (e.g. highlight colour) will be saved even if the match-set group is toggled off and then on again.
- If you delete the match-set group using the 'Edit Groups' dialog, all of its settings will be lost; you will get the default settings again the next time you enable the match-set group. To avoid this, disable it by toggling it off using the 'Toggle match set' menu option (or 'g' shortcut key) rather than by deleting it in the Groups dialog.

Editing groups:

To edit a group, right-click and select 'Edit Groups', or use the Ctrl-G shortcut key.

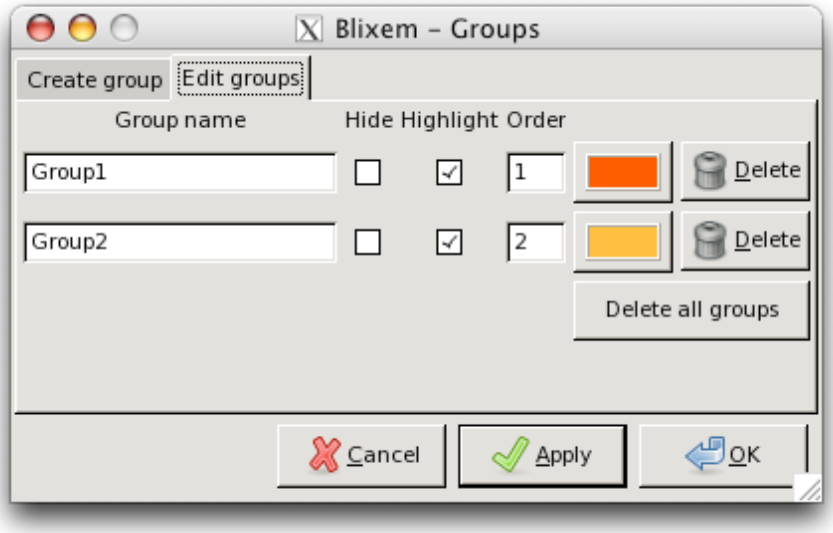


Figure 11: Groups dialog - edit groups

You can change the following properties for a group. Click on Apply or OK to apply the changes.

Name	You can specify a more meaningful name to help you identify the group.
Hide	Tick this box to hide the alignments in the alignment lists.

Highlight	Tick this box to highlight the alignments.
Colour	The colour the group will be highlighted in, if 'Highlight' is enabled. The default colour for all groups is orange, so you may wish to change this if you want different groups to be highlighted in different colours.
Order	When sorting by Group, alignments in a group with a lower order number will appear before those with a higher order number (or vice versa if sort order is inverted). Alignments in a group will appear before alignments that are not in a group.

To delete a group, click one of the following buttons. This will have an immediate effect (i.e. you don't have to click 'Apply').

- To delete a single group, click on the 'Delete' button next to the group you wish to delete.
- To delete all groups, click on the 'Delete all groups' button.

Running dotter

- To start Dotter from within Blixem, or to edit the parameters for running Dotter, right-click and select 'Dotter' or use the Ctrl-D keyboard shortcut. The Dotter dialog will pop up.

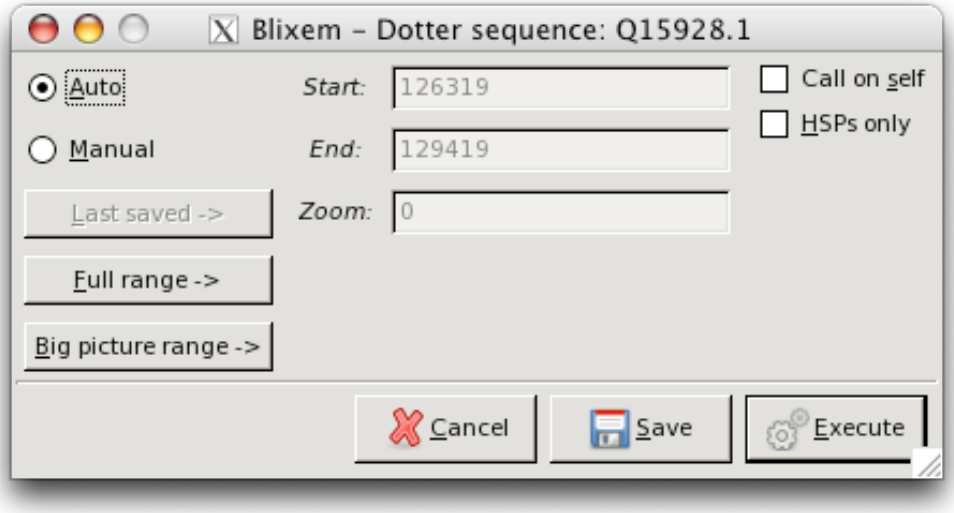


Figure 12: Dotter dialog

- Select the sequence you wish to run Dotter on before or after opening the dialog. The selected sequence name will be shown at the top of the dialog.
- Alternatively, if you just wish to edit the settings, you do not need to select a sequence.
- To run Dotter with the default (automatic) parameters, just hit RETURN, or click the 'Execute' button.
- To enter custom parameters, select the 'Manual' radio button and enter the values in the 'Start' and 'End' boxes.
- To save the parameters without running Dotter, click Save and then Cancel'.
- To save the parameters and run Dotter, click 'Execute'.

- To revert to the last-saved manual parameters, click the 'Last saved' button.
- To revert back to automatic parameters, click the 'Auto' radio button. The coordinates in the Start and End box will be recalculated for the currently-selected sequence.

Reference sequence versus itself

To run Dotter on the reference sequence versus itself, select the ‘Call on self’ tick box in the Dotter dialog and then click ‘Execute’. This can be useful to analyse internal repeats etc. (see the Dotter manual for more information).

Dotter HSPs only

This starts Dotter in HSP (High-Scoring Pair) mode. See the Dotter manual for more information.

Settings

The settings menu can be accessed by right-clicking and selecting Settings, or by the shortcut Ctrl-S.

Features

Highlight variations

When this option is enabled, bases in the reference sequence that have know variations (such as SNPs, insertions, deletions etc.) are highlighted in the reference sequence (nucleotide) header. If the 'Show variations track' sub-option is also enabled, then an additional line is shown above the nucleotide header showing the alternative bases for each variation. Note that the Variations track can be quickly enabled or disabled by double-clicking the nucleotide header.

Show polyA tails

When this option is enabled, polyA tails are shown and highlighted in the alignment lists and polyA signals are highlighted in the reference sequence (nucleotide) header. If the sub-option 'Selected sequences only' is enabled, polyA features will only be shown for the currently selected sequences.

Display options

Show Unaligned Sequence

When this option is enabled, any additional, unaligned portions of the match sequences are displayed at the start and end of the alignments. If the 'Limit to' sub-option is also enabled, you can specify the maximum number of additional bases to display. If the 'Selected sequences only' sub-option is enabled, only the currently selected sequence(s) will display unaligned portions of sequence.

Show Splice Sites

When this option is enabled, splice sites are highlighted in the reference sequence (nucleotide) header for the currently-selected sequence(s). The two bases from the adjacent introns are highlighted in green if they are canonical or red if they are non-canonical.

Highlight Differences

When this option is enabled, matching bases are blanked out and mismatches are highlighted, making it easier to see where alignments differ from the reference sequence.

Squash Matches

This groups multiple alignments from the same sequence together into the same row in the detail view, rather than showing them on separate rows.

Invert Sort Order:

Reverse the default sort order. (Note that some columns sort ascending by default (e.g. name, start, end) and some sort descending (score and ID). This option reverses that sort order.)

General settings

Font

Allows you to change the font that is used to display alignments in the detail-view. Note that you must select a monospace font; otherwise matches will not be shown aligned correctly. Blixem will warn you if the font you have selected is not monospace.

Fetch mode

Allows you to change the program used to fetch sequence EMBL entries. (Currently only available to authorised users within the Sanger Institute).

Columns

Load optional data

Click this button to load optional data from EMBL entries (currently only applicable to authorised users within the Sanger Institute). Note that this operation can take a long time if there are many sequences. The button will be greyed out once optional data has been loaded.

Column visibility

Tick/un-tick the check-marks to show/hide individual columns. Adjust the column width by entering the new width in the text box in pixels. Note that if you enter a zero width then the column will be hidden, regardless of whether the check-mark is ticked or not. Greyed-out columns are optional-data columns, and will only become available once optional data has been loaded.

Grid properties

%ID per cell

Use this to change the vertical scale of the grid; a smaller value means the grid will be more spaced out, a larger value means the grid will be more compact.

Max %ID

Defines the maximum cut-off value for the %ID scale.

Min %ID

Defines the minimum cut-off value for the %ID scale.

Appearance

Use print colours

Select this option to make Blixem use grey-scale colours, suitable for printing.

Display colours

Change any of Blixem's custom display colours, such as the colour aligned bases are shown in or the colour stop codons are highlighted in etc. There are four colours for each item:

- Normal: this is the standard display colour;
- Normal (selected): this is the colour used when the item is selected (if applicable). Typically one would use a slightly darker or lighter shade of the Normal colour for this, so that the item does not look radically different when it is selected;
- Print: this is the standard colour used when the 'Use print colours' option is enabled;
- Print (selected): this is the colour used when 'Use print colours' is enabled and the item is selected.

Key

In the detail view, the following colours and symbols have the following meanings:

Alignment list header	Yellow background	Reference sequence
Alignment list	Cyan background	Identical residues
Alignment list	Violet background	Conserved residues
Alignment list	Grey background	Mismatch
Alignment list	'.' with grey background	Deletion
Alignment list	Yellow vertical line	Insertion
Alignment list	Thin blue vertical line	Boundary of an exon
Nucleotide header (protein mode)	Sky-blue background	The three nucleotides for the currently-selected codon; darker blue indicates the nucleotide whose coordinate is displayed in the feedback box
Alignment list header (protein mode)	Pale red background	STOP codon
Alignment list header (protein mode)	Green background	MET codon

Keyboard shortcuts

Ctrl-Q	Quit
Ctrl-H	Help
Ctrl-P	Print
Ctrl-S	Edit settings
V	Show/hide sections of the display
Shift-Ctrl-G	Create group
Ctrl-G	Edit groups (or create a group if none currently exist)
Ctrl-A	Select all sequences in the current list
Shift-Ctrl-A	Deselect all sequences
Ctrl-D	Dotter
Left-arrow	Move coordinate section one index to the left[2]
Right-arrow	Move coordinate section one index to the right[2]
Shift-Left	Same as Left, but in protein mode it scrolls by a single nucleotide
Shift-Right	Same as Right, but in protein mode it scrolls by a single nucleotide
Ctrl-Left	Scroll to the start/end of the previous alignment[3]
Ctrl-Right	Scroll to the start/end of the next alignment[3]
Up-arrow	Move row selection up
Down-arrow	Move row selection down
Home	Scroll to the start of the display
End	Scroll to the end of the display
Ctrl-Home	Scroll to the start of the first alignment[3]
Ctrl-End	Scroll to the end of the last alignment[3]
=	Zoom in detail view
-	Zoom out detail view
Ctrl-=	Zoom in big picture
Ctrl--	Zoom out big picture
Shift-Ctrl--	Zoom out big picture to view the whole reference sequence
,	Scroll left one coordinate
.	Scroll right one coordinate
P	Go to position
T	Toggle the active strand
G	Toggle the 'match set' Group
1	Toggles visibility of the 1st alignment list
2	Toggles visibility of the 2nd alignment list
3	Toggles visibility of the 3rd alignment list (protein mode only)
Ctrl-1	Toggles visibility of the 1st big picture grid
Ctrl-2	Toggles visibility of the 2nd big picture grid
Shift-Ctrl-1	Toggles visibility of the 1st exon view
Shift-Ctrl-2	Toggles visibility of the 2nd exon view

[2] **Only applicable if a coordinate is currently selected; middle-click a coordinate to select it.**

[3] **Limited to just the selected sequences, if any are selected; otherwise, acts on all sequences.**

Dotter User Manual

Written by Gemma Barson
(gb10@sanger.ac.uk)

Wellcome Trust Sanger Institute
18 January 2011

Dotter

This manual explains how to configure, run and use Dotter. Dotter is a graphical dot-plot program for detailed comparison of two sequences. Every residue in one sequence is compared to every residue in the other sequence. The first sequence runs along the x-axis and the second sequence along the y-axis. In regions where the two sequences are similar to each other, a row of high scores will run diagonally across the dot matrix.

Dotter is maintained by the Wellcome Trust Sanger Institute and is available as part of the SeqTools package. The software can be downloaded from the Sanger Institute's website: http://www.sanger.ac.uk.

Getting Started

Running Dotter

As a minimum, Dotter takes the following required arguments:

```
dotter <horizontal_sequence> <vertical_sequence>
```

where `<horizontal_sequence>` and `<vertical_sequence>` are the path names of FASTA files containing the two input sequences.

Dotter will assume that the sequences both start at coordinate 1 unless you use the `-q` and `-s` arguments to set an offset for the query (horizontal) and subject (vertical) sequences respectively.

Run '`dotter`' without any arguments to see further usage information.

Sequence versus itself

Dotter can be run on a sequence versus itself. This can be useful to analyse internal repeats. If you're comparing a sequence against itself, you'll notice that the main diagonal scores maximally, since it's the 100% perfect self-match.

Input files

The sequence input files are in FASTA format. Comparisons are allowed between two nucleotide sequences, two protein sequences, or one nucleotide and one protein sequence – note that when comparing a nucleotide and a protein sequence, the nucleotide sequence must be passed first (i.e. as the horizontal sequence).

Additional features can be passed to Dotter in a GFF file using the `-f` argument. Relevant features include alignments, which can be viewed using Dotter's HSP mode, and transcripts, which are shown at the bottom of the Dotter window.

FASTA file:

A FASTA file has a header line that starts with '>' and contains the sequence name. The next line contains the start of the sequence data. The sequence data can be on a single line or separated by newlines; it is usually separated by newlines every 50 characters to aid readability.

```
>chr4-04_210623-364887
tcttgtttctgtaggagaggccatctccatcagctataaccaaaaaaaaa
acaaaaaactcctctttttgacaagtttgtaaagcctgtccatctgggtc
tataataatcctccaggccctatgccactcctctttattcagccagttca
...
```

GFF file:

Dotter uses the GFF version 3 file format. In this section we give a very brief description of this file format; see http://www.sequenceontology.org/gff3.shtml for a full description.

The GFF file should start with the following two comment lines. (Additional comments can be included but may be ignored.)

```
##gff-version 3
##sequence-region chr4-04_210623-364887 44144 154265
```

Each subsequent line defines a feature. A feature line must have the following 8 tab-separated columns:

```
reference_sequence_name       source  type    start   end     score   strand  phase
```

An optional 9th column defines any tags (separated by semi-colons). Dotter supports the following GFF tags. (Additional tags can be supplied but may be ignored.)

Target (required for alignments)
Gap (required for gapped alignments)
ID (required for parent features)
Name (required for transcripts and SNPs)
Parent (required for child features)

Transcripts

Note that exons should have a Parent transcript defined, and the Name tag should be set in the parent rather than the child exons. Note that Dotter *will* recognise exons that do not have a Parent tag if they have a Name tag instead, but they may not get grouped correctly with other exons from the same transcript.

Typically, one defines the parent transcript, the exons, and the CDS regions; Dotter will then calculate the missing components (in this case, the UTR regions and the introns). Dotter will recognise other combinations of inputs, and will always calculate the missing components as long as enough information is provided.

Sample GFF file

A sample GFF file may look like this ('…' denotes that text has been omitted).

```
##gff-version 3
##sequence-region chr4-04_210623-364887 44144 154265
chr4-04_210623-364887  EST_Human      nucleotide_match      79195   79311   95.000000
       -       .       Target=DA692754.1 287 403 +;percentID=90.6;sequence=GATCTGGC...
chr4-04_210623-364887  EST_Human      nucleotide_match      79195   79323   121.000000
       +       .       Target=AI095103.1 326 454 +;percentID=96.9;sequence=TTTAAATT...
chr4-04_210623-364887  ensembl_variation     deletion       80798   80799   .       +
       .
       Name=rs60725655;url=http%3A%2F%2Fwww.ensembl.org%2FHomo_sapiens%2FVariation%2FSumm
ary%3Fv%3Drs60725655;variant_sequence=AA/-;
chr4-04_210623-364887  Augustus       mRNA    119534  119941  .       -       .
       ID=transcript21;Name=AUGUSTUS00000051712
```

```
chr4-04_210623-364887  Augustus       exon    119534  119941  .       -       .
       Parent=transcript21
chr4-04_210623-364887  Augustus       CDS     119534  119941  .       -       0
       Parent=transcript21
```

The Dotter Windows

The dot-plot window

The main Dotter window contains the dot-matrix plot. It also shows any exons for the sequences along the bottom of the window (for the horizontal sequence; or along the right-hand-side for the vertical sequence).

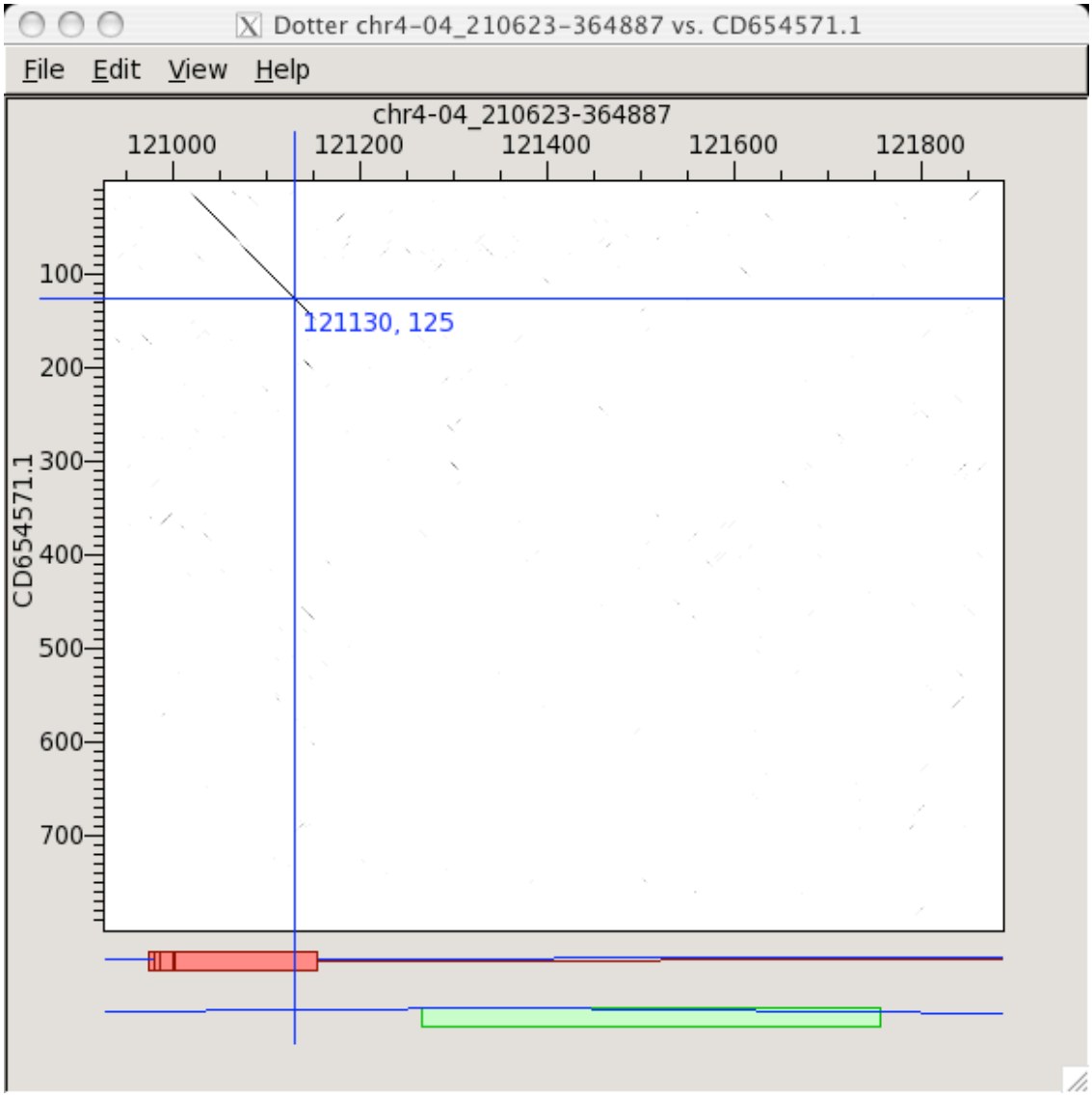


Figure 13: The main window

Cross-hair

The blue cross-hair shows the coordinates at a particular position. It can be moved by clicking/dragging with the left mouse button, or by using the following keyboard

shortcuts:

Left-arrow **Right-arrow**	Move one dot left/right along the horizontal sequence.
Shift-Left **Shift-Right**	The same as Left/Right, but for protein sequences this moves by a single nucleotide coordinate rather than a whole dot/amino-acid.
Up-arrow **Down-arrow**	Move one dot up/down along the vertical sequence.
Shift-Up **Shift-Down**	The same as Up/Down, but for protein sequences this moves by a single nucleotide coordinate rather than a whole dot/amino-acid.
, .	Move diagonally up-left or down-right. Useful for moving along an alignment.
[ ]	Move diagonally down-left or up-right. Useful for moving along an alignment.

Zoom in with a child Dotter

You can open a new child Dotter on a particular region from the current Dotter window. Middle-click and drag the mouse to select the region to open the new Dotter on.

The alignment tool

The alignment tool shows the portions of the two sequences at the current cross-hair position. The sequences will move to remain centred on the cross-hair coordinates when the cross-hair is moved. The same shortcut keys for moving the cross-hair can be used in this window.

Aligning matches are highlighted and colour-coded according to whether they are an exact or conserved match (cyan for exact, violet for conserved).

In nucleotide->nucleotide mode, both strands of the horizontal sequence are shown in the alignment tool. In nucleotide->protein mode, all three reading frames of the horizontal sequence are shown, and the best match out of the three frames determines the highlight colour for the bases in the vertical sequence.

If closed or hidden, the alignment tool can be shown with the 'Ctrl-A' shortcut or by selecting the 'Alignment tool' option under the 'View' menu.

Figure 14: Alignment tool - nucleotide->nucleotide mode

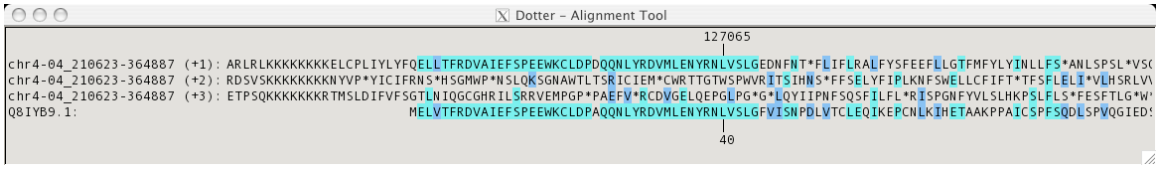
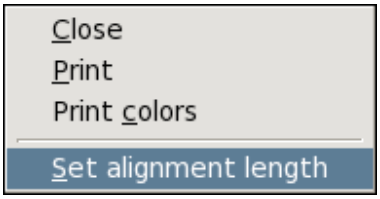


Figure 15: Alignment tool - nucleotide->protein mode

Alignment tool menu

Right-clicking in the alignment tool brings up a context menu. The 'Set alignment length' option allows you to specify how long a portion of the sequences should be shown in the alignment tool.



Greyramp tool

This tool controls the threshold and contrast of the the dot-plot image. To improve visualization, little peaks (noise) can be nullified by a minimum cut-off. Similarly, significant peaks above a certain score can be saturated by a maximum cut-off.

Drag the square handle and the arrows to change the threshold and contrast. The 'Swap' button swaps the positions of the top and bottom arrows, inverting the colours. The 'Undo' button undoes the effect of the last drag.

If closed or hidden, the greyramp tool can be shown with the 'Ctrl-G' shortcut or by selecting the 'Greyramp tool' option under the 'View' menu.

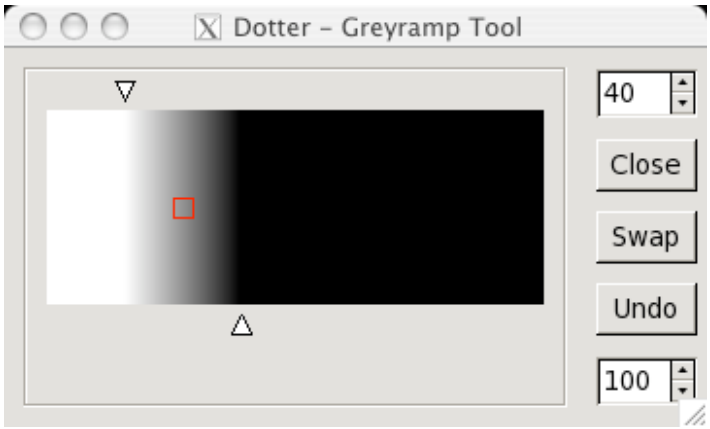
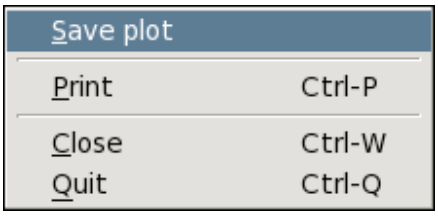


Figure 16: Greyramp tool

Main menu

The main menu can be accessed via the menu-bar at the top of the dot-plot window or by right-clicking in the dot-plot window.



File menu

Save plot: Save the current dot-plot. It can be re-loaded by calling Dotter from the command line using the -l argument. Note that you will need to call Dotter with the same portion of each sequence that was originally passed to Dotter in order for the alignment tool to function correctly when you load the dot-plot.

Print: Print the current dot-plot.

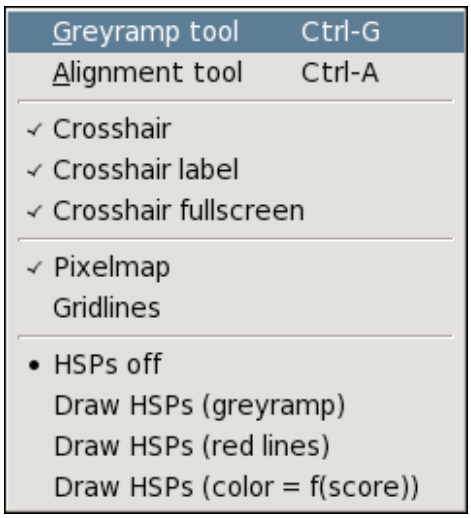
Close: Close the current Dotter window. Also closes the associated alignment and greyramp tool, but does not close any other Dotter windows.

Quit: Close the current Dotter window and all associated Dotters as well (including any child or parent Dotters). If you just wish to close the current Dotter, then use the 'Close' menu option instead.

Edit menu

Settings: Show the 'Settings' dialog.

View menu



Greyramp tool: Show the greyramp tool.

Alignment tool: Show the alignment tool.

Crosshair: Toggle visibility of the cross-hair

Crosshair label: Toggle visibility of the cross-hair label (only has an effect if the cross-hair is visible).

Crosshair fullscreen: Toggle whether the cross-hair is shown to its full extents or is clipped to just the dot-plot area.

Pixelmap: Toggle visibility of the grey-scale dot-plot image.

Gridlines: Toggle visibility of gridlines.

HSPs off: Select this option to turn HSP (High Scoring Pair) mode off.

Draw HSPs (greyramp): Select this option to view HSPs in grey-scale mode. In this mode, the HSPs (High Scoring Pairs) are drawn in a shade of grey that is determined by their score. The greyramp tool can be used to adjust the thresholds and contrast of the HSP image. This mode *replaces* the standard dot-plot image.

Draw HSPs (red lines): Select this option to view all HSPs as red lines. This mode can be used in conjunction with the standard dot-plot image: HSPs are drawn over the top.

Draw HSPs (color=f(score)): Select this option to view HSPs as solid lines, whose colour depends on their score. This mode can be used in conjunction with the standard dot-plot image: HSPs are drawn over the top.

Help menu

Help: Show the 'Help' dialog.

About: Show the 'About' dialog.

Settings

The settings menu can be accessed by selecting the 'Settings' option on the 'Edit' menu, or by pressing the 'Ctrl-S' shortcut key.

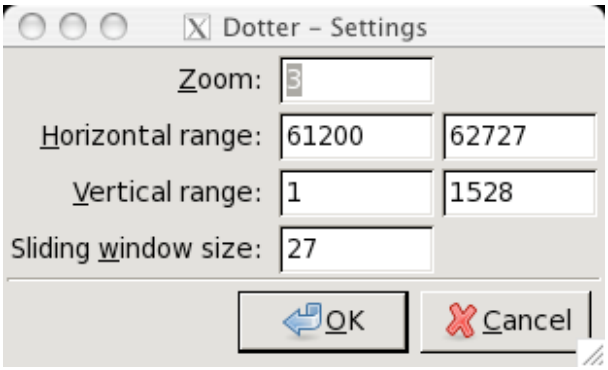


Figure 17: The Settings menu

Zoom

Specify the zoom factor. The factor is an inverse: a zoom factor of 3 will zoom *out* by a factor of 3, i.e. the window will shrink to 1/3 of its full size. A zoom factor of 1 will show the window at full size. A factor of less than 1 (e.g. 0.5) can be set in order to zoom in, but this will result in a stretched dot-plot so is not recommended.

Horizontal range

Set the range of the horizontal sequence. The maximum range possible is the range that was originally passed to Dotter – the range you enter will be trimmed if you enter out-of-range values.

Note that this causes the matrix to be recalculated, so if it took a long time to calculate in the first place, stay away from this menu item!

Vertical range

Set the range of the vertical sequence. The maximum range possible is the range that was originally passed to Dotter – the range you enter will be trimmed if you enter out-of-range values.

Note that this causes the matrix to be recalculated, so if it took a long time to calculate in the first place, stay away from this menu item!

Sliding window size

To make the score matrix more intelligible, the pairwise scores are averaged over a sliding window that runs diagonally. This option allows you to edit the size of the sliding window. There's normally no need to change this.

Note that this causes the matrix to be recalculated, so if it took a long time to calculate in the first place, stay away from this menu item!

Keyboard shortcuts

Left-arrow **Right-arrow**	Move the cross-hair one dot left/right along the horizontal sequence.
Shift-Left **Shift-Right**	The same as Left/Right, but for protein sequences this moves by a single nucleotide coordinate rather than a whole dot/amino-acid.
Up-arrow **Down-arrow**	Move the cross-hair one dot up/down along the vertical sequence.
Shift-Up **Shift-Down**	The same as Up/Down, but for protein sequences this moves by a single nucleotide coordinate rather than a whole dot/amino-acid.
, **.**	Move diagonally up-left or down-right. Useful for moving along an alignment.
[**]**	Move diagonally down-left or up-right. Useful for moving along an alignment.
Ctrl-W	Close the current window. If this is a dot-plot window, it also closes the associated alignment and greyramp tool.
Ctrl-Q	Quit Dotter. Also quits any associated Dotters, i.e. any child or parent Dotters.
Ctrl-S	Open the Settings dialog.
Ctrl-H	Open the Help dialog.
Ctrl-A	Show the alignment tool.
Ctrl-G	Show the greyramp tool.
Ctrl-D	Show the main dot-plot window.

Annotation resources

AspicDB – useful analysis of splice junctions http://t.caspur.it/ASPicDB/

CCDS http://www.ncbi.nlm.nih.gov/projects/CCDS/CcdsBrowse.cgi

Ensembl genome browser http://www.ensembl.org/index.html

Entrez Gene for nucleotide and protein sequence, cloning, gene information etc http://www.ncbi.nlm.nih.gov/sites/gquery

HORDE database for olfactory receptors http://genome.weizmann.ac.il/horde/

Swiss Institute of Bioinformatics has many tools for analysing nucleotide and protein sequences http://www.expasy.ch/

UCSC genome browser http://genome.ucsc.edu/cgi-bin/hgGateway

UniProt has protein sequence information http://www.uniprot.org/

Vertebrate Genome Annotation Browser for manual annotation http://vega.sanger.ac.uk/index.html